Deep Domain Adaptation Based Multi-Spectral Salient Object Detection

Shaoyue Song, Zhenjiang Miao, *Member, IEEE*, Hongkai Yu, *Member, IEEE*, Jianwu Fang, *Member, IEEE*, Kang Zheng, Cong Ma, and Song Wang, *Senior Member, IEEE*

***Abstract*—Salient Object Detection (SOD) plays an important role in many image-related multimedia applications. Although there are many existing research works about the salient object detection in traditional RGB (visible-light spectrum) 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 mainly model this research problem as a deep learning based domain adaptation from the traditional RGB image data (source domain) to the multi-spectral data (target domain), and an adversarial deep domain adaptation model is proposed. We first collect and will publicize a large multi-spectral dataset, RGBN-SOD dataset, including 780 synchronized RGB and NIR image pairs for the multi-spectral SOD problem in the simple and complex situations. Intensive experimental results show the effectiveness and accuracy of the proposed deep domain adaptation for the multi-spectral SOD. Besides, due to the absence of research on the field of multi-spectral co-saliency detection, we also collect 200 synchronized RGB and NIR image pairs in addition to explore the multi-spectral co-saliency detection.**

***Index Terms*—Domain adaptation, multi-spectral, salient object detection.**

Manuscript received April 30, 2020; revised September 5, 2020 and November 10, 2020; accepted December 12, 2020. Date of publication December 25, 2020; date of current version January 21, 2022. This work was supported in part by the NSFC 61672089, 61703436, 61572064, 61273274, CELFA; NSFC-61672376, NSFC-U 1803264; and NSFC 61806022, in part by Fundamental Research Funds for the Central Universities, CHD (No.300102320202), and in part by AWS Cloud Credits for Research Award. This article was presented in part at the Conference on Artificial Intelligence, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Engin Erzin. *(Corresponding authors: Zhenjiang Miao; Hongkai Yu.)*

Shaoyue Song, Zhenjiang Miao, and Cong Ma are with the Institute of Information Science and Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: 14112060@bjtu.edu.cn; zjmiao@bjtu.edu.cn; 13112063@bjtu.edu.cn).

Hongkai Yu is with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH 44115 USA (e-mail: h.yu19@csuohio.edu).

Jianwu Fang is with the College of Transportation Engineering, Chang'an University, Xi'an 710100, China (e-mail: j.w.fangit@gmail.com).

Kang Zheng and Song Wang are with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208 USA (e-mail: zheng37@email.sc.edu; songwang@cec.sc.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3046868.

Digital Object Identifier 10.1109/TMM.2020.3046868

I. Introduction

SALIENT object detection (SOD) which aims at finding out the salient objects in a given image is quite helpful to discover the objects and well understand the image scene, so the SOD techniques could benefit many image-related multimedia applications, such as image scene understanding [2], image segmentation [3], [4], object tracking [5], common object discovery [6], [7], etc. There are many existing research works about the SOD in RGB images such as [8]–[12], 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. In this paper, we discuss the new multi-spectral SOD problem as shown in Fig. 1. Traditional RGB SOD problem takes one single RGB image in the visible-light spectrum as the input, while the multi-spectral SOD problem uses the multi-spectral images as the input. We expect that the multi-spectral images could improve the SOD performance in some complex situations. Situations like shadow effect, similar appearance between background and foreground, strong or insufficient illumination, as shown in Fig. 2, are taken as complex situation in this paper. In complex situations, it is usually difficult to find out the salient object. Some related researches also show that it is necessary to study the applications under complex situations, such as the indistinguishable object appearance [13], the poor quality images obtained in real-world changing environments or adverse weather conditions [14], [15] and so on.

In order to better handle the complex situations, additional modality information could be introduced to help the SOD, like the mid-wave infrared images [13], thermal infrared images [16], depth images [17], [18], etc. Different kinds of image modalities can provide different unique contributions for SOD. Near-infrared (NIR) spectrum is one of the image modalities which has shown successes in many image processing and multimedia tasks. The NIR image is often used to help the RGB image tasks such as the low-light image enhancement [19], image restoration [20], image dehazing [21], [22], image denoising [23], robust scene category recognition [24], face recognition robust to illumination variations [25], image quality and context improvement to the changeable weather [26], etc. For example, as shown in Fig. 2, the RGB images might show low discriminative contrast in complex situations, while the synchronized NIR images might display a better contrast to human vision systems. Studies on the face recognition [25], [27] show that

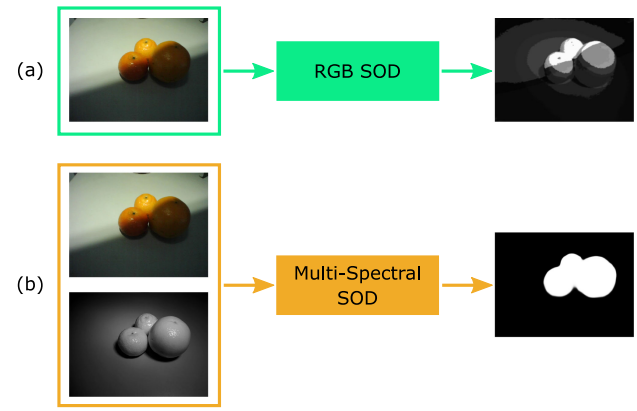


Fig. 1. Illustration of the multi-spectral SOD problem. Traditional RGB SOD problem takes one single RGB image in the visible-light spectrum as the input shown in (a), while the multi-spectral SOD problem uses the multi-spectral images as the input shown in (b).

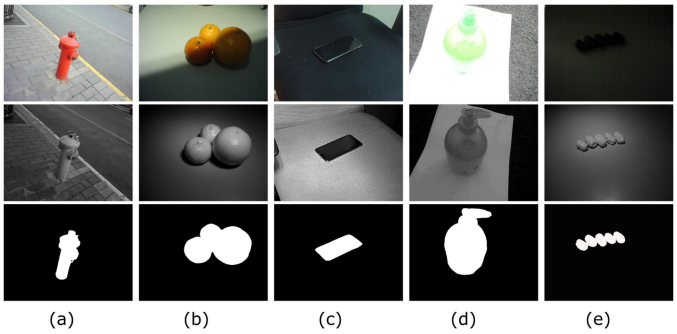


Fig. 2. Sample images from the collected multi-spectral dataset in simple and complex situations: (a) simple/normal situation, (b) shadow effect, (c) similar appearance between background and foreground, (d) strong illumination, and (e) insufficient illumination. From top to bottom: RGB image, synchronized NIR image, annotated ground truth for SOD. We can see that NIR images could improve the contrast of the salient object and also keep the object details information in complex situations.

the NIR image is less sensitive to the variations of visible illuminations. The study in [28] indicates that using visible RGB and NIR image pair can improve the visual quality on those image influenced by changeable weather like haze, fog, smoke and so on. The study in [26] shows that RGB images contain more color information while NIR images contain rich sharp edge information. Taking both of the RGB and NIR information into consideration might improve the SOD performance under the degraded situations. In many real-world image related applications like robotics [29], [30], autonomous vehicles [31], and video surveillance [32], 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, this paper explores 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 datasets (such as the well-labelled MSRA-B [33], DUTS [34], HKU-IS [35]) could provide references for the multi-spectral SOD problem. We model the SOD problem with additional NIR information as an unsupervised domain adaptation problem in deep learning. Domain Adaptation (DA) aims to use the information from both source and target domains to reduce the domain discrepancy [36]–[39]. Recently, DA is popular in the research field of autonomous driving [40], medical imaging segmentation [41], etc. We adopt a deep learning based adversarial domain adaptation method to reduce the domain discrepancy for the proposed multi-spectral SOD problem.

Existing datasets using NIR images [42], [43] 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, named as RGBN-SOD 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 in deep learning 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. In addition, we also collect a new dataset (200 RGB-NIR image pairs) for the multi-spectral co-salient object detection problem.
3) We propose a new deep learning 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).

A preliminary version of our work has already been published in our previous conference proceeding [1]. Compared with [1], this journal paper adds more detailed explanations in the introduction and related work, a new experiment on another multi-spectral SOD dataset [42], a new experimental comparison with more state-of-the-art SOD models, new experiments on the larger-scale datasets, and the failure cases. In addition, we also explore the problem of the multi-spectral co-salient object detection in this journal paper.

II. 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. SOD as a fundamental computer vision task has been one of the popular research fields for many years. The traditional methods like RC [8], LRK [44], CWS [45], FT [46] usually concentrate on some specific low-level features and certain prior information like connectivity prior [47], background prior [48]. Recently, by the powerful representation of deep learning based methods, the SOD task performance is improved a lot [49].

Recently, many deep learning based methods [9], [50]–[57] achieve good performance in the SOD research. For instance,

in [51], SOD problem is solved in the pixel level instead of the patch level with the proposed end-to-end deep contrast network. In [50], better methods are developed for fusing multi-scale information for SOD. Recently, many researches like [53], [57], [58] focus more on the edge/boundary information to help obtain more accurate salient object detection results. Attention mechanism [56], [58] is also introduced into the design of SOD models to enhance the representation ability of the SOD models. In [59], [60], salient objects are detected from fixation maps by a deep hierarchy of convolutional LSTMs. Besides, there are also some researches focusing more on the speed and the computation cost of the SOD models. For example, in [52], a fast and accurate salient object detection method is proposed by directly utilizing the generated saliency maps to refine the features of the backbone network and in [61], [62] a novel dynamic weight decay scheme is proposed to reduce the representation redundancy.

Most of deep learning based SOD methods are based on RGB images only and have not considered the multi-spectral cues. There are many well-labelled SOD datasets of RGB images such as MSRA-B [33], DUTS [34], HKU-IS [35], etc. In this paper, we suppose that the RGB SOD datasets can provide references and guides for the multi-spectral SOD problem.

RGBD image SOD: The RGB-Depth (RGBD) based SOD task which takes use of the depth image in SOD task [63]–[67] is similar to our multi-spectral SOD task, but there are many differences between them. The physical meaning of these two kinds of image modalities is different. Near-infrared image in this paper refer to images containing the information of the object under the wavelength of the near-infrared band (850 nm), while the depth image usually reflects the distance mapping of a scene. The NIR image contains more detailed scene and object information, while the depth image loses many details by only considering the distance cue. Besides, the NIR image can distinguish similar-appearance objects of different materials [68] even when their depths are the same or very similar.

Mutli-spectral SOD and related datasets: No matter the images captured under the outdoor or indoor environments will be easily suffered from the image degradation by the changing illumination, bad weather conditions, complex situations, as shown in Fig. 2. Directly applying the SOD methods developed based on regular images to the degraded images might fail in extracting the salient object, so it is necessary to find an efficient way to improve the SOD performance in these complex situations. The SOD problem in complex situations might be improved via other kinds of image modalities providing extra and unique features. Some researches use other image modalities to help the salient object detection for the RGB images, for example: [69] uses the flash and no-flash image pairs to solve the SOD problem; [70] collects a RGB and thermal infrared dataset to study the SOD problem; [71] solves the foggy image SOD problem via the fusion of spatial and frequency domains. Because near-infrared images contain similar structures as the visible color images [20] and also provide more edge and detail information about the object and image scene, this paper tries to include multi-spectral modalities (i.e., RGB and NIR images) for the SOD problem.

The most related datasets to our SOD problem are the existing multi-spectral datasets including NIR images [42], [43]. 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 and do not consider the complex situations. In this paper, we first collect a large multi-spectral dataset including 780 synchronized RGB and NIR image pairs for the multi-spectral SOD problem in both simple and complex situations. Besides, we also collected 200 synchronized RGB and NIR image pairs for the multi-spectral co-saliency detection problem.

Deep domain adaptation: In deep learning researches, domain adaptation is to reduce the data distribution discrepancy in the source and target domains so as to improve the generalization ability of the deep learning model [38]. This paper focuses on the adversarial domain adaptation in deep learning for the multi-spectral SOD problem due to its advanced performance recently. In the research of adversarial domain adaptation, generative adversarial learning [72] could be used to reduce the domain shifts across different domains. Typically, a generator and a discriminator are trained against each other [73], [74]. 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 [73]–[75]. A new deep learning based 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) is proposed.

III. Multi-Spectral SOD Dataset

We firstly collected a new dataset named as RGBN-SOD 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) in the RGBN-SOD dataset.

1) 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 mainly select 4 kinds of situation as complex situations and 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 RGBN-SOD dataset is shown in Fig. 3 (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. 2. 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.

2) 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,

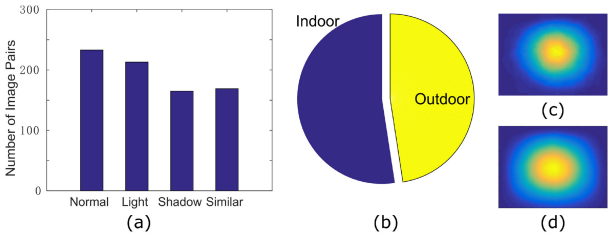


Fig. 3. Statistics for the proposed multi-spectral SOD dataset RGBN-SOD: (a) distribution under simple/normal and complex situations, (b) distribution under indoor and outdoor scenes, (c) average ground-truth on the proposed dataset, (d) average ground-truth on MSRA-B dataset [33].

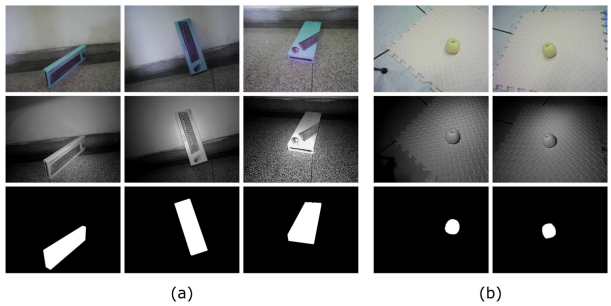


Fig. 4. Two groups of sample image pairs in the collected co-RGBN dataset for co-saliency detection. Each group has multiple RGB-NIR image pairs containing common salient objects. Top: RGB image; Middle: NIR image; and Bottom: ground truth.

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 640×480 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 RGBN-SOD is shown in Fig. 3 (c).

In addition, we then collected a new dataset for the multi-spectral co-salient object detection problem [18], named as co-RGBN dataset in this paper. This co-RGBN dataset containing 200 RGB-NIR image pairs is divided into 58 groups. In each group, there are multiple RGB-NIR image pairs containing common salient objects. Sample images of the groups of the co-RGBN dataset are shown in Fig. 4. The target is to detect the common salient object(s) in each group of multiple images.

IV. Methodology

In this section, we 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.

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

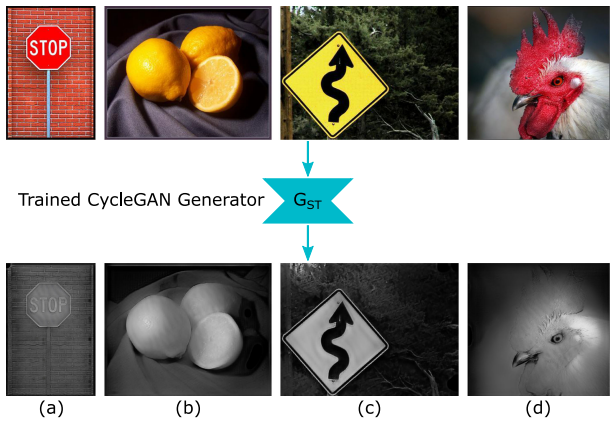


Fig. 5. Examples of pseudo-NIR image generation. Top row: RGB images and bottom row: generated pseudo-NIR images. The CycleGAN generator G_{ST} trained between the RGB images of the MSRA-B dataset and the NIR images of RGBN-SOD dataset is used to generate the corresponding pseudo-NIR images for the given source domain RGB images.

target-domain image pairs X_T^{rgb} and X_T^{nir} without the pixel-wise 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. 6.

A. 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 [76], this translation is an unpaired image-to-image transfer, which can be achieved by the advanced CycleGAN [77].

CycleGAN [77] 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 RGBN-SOD dataset, following the network structure and setup in CycleGAN [77], 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. 5. With the help of CycleGAN based pseudo-NIR image generation, the cross-domain data distribution (mainly between different datasets) discrepancy is somewhat reduced. Experimental

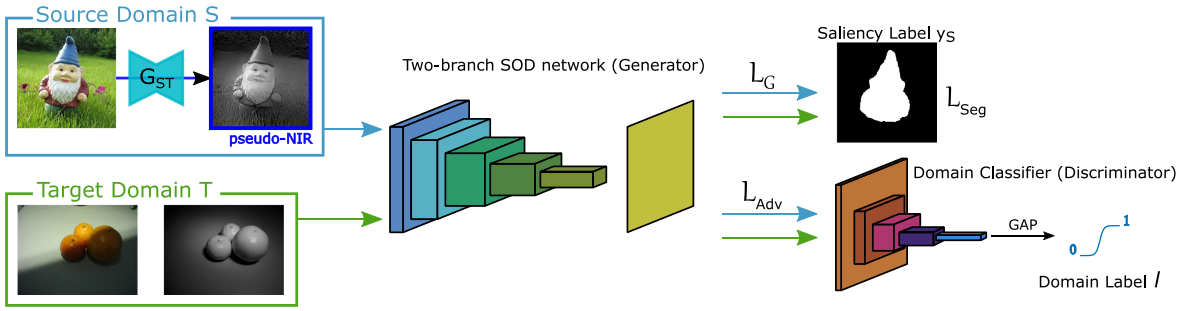


Fig. 6. Framework of the proposed adversarial domain adaptation method for the unsupervised multi-spectral SOD. It consists of a trained CycleGAN generator G_{ST}, a two-branch SOD network (Generator) and a domain classifier (Discriminator). The source domain S is an existing RGB SOD dataset like MSRA-B [76] with the pixelwise ground-truth labels and the target domain T is the proposed multi-spectral SOD dataset without the pixelwise ground-truth labels.

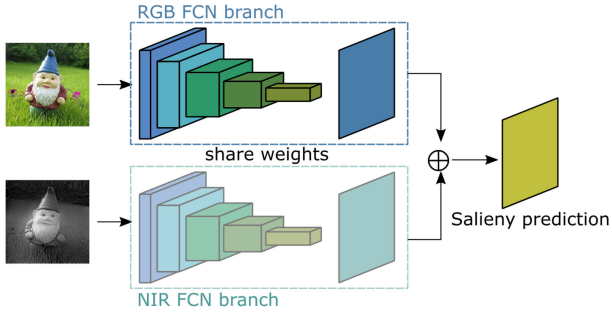


Fig. 7. Proposed two-branch SOD network with the RGB and NIR (or pseudo-NIR) image pair as the input.

results also display the effectiveness of the pseudo-NIR image generation in domain shift reduction.

B. 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) [78] 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 [51].

As shown in Fig. 7, 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 pixelwise addition. In our experiment, we adopt VGG16 [79] as our backbone network for FCN, and other FCN models can also be applied to our proposed framework. The proposed two-branch SOD network is simple but efficient to capture the united multi-spectral cues from RGB and NIR images.

C. 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. 6. 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 [80] as a reference. D has five stacked strided convolutional layers with 3×3 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 [80], 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 pixelwise 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. As showing in Fig. 6, the pseudo-NIR image of X_S is generated by the corresponding RGB image

and the trained G_{ST}. y_s is the two-class pixelwise ground-truth map of salient and non-salient classes. Like [72], [74], 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:

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

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 X_T is a target-domain RGB-NIR image pair, λ_1 and λ_2 are the balance weights and we set them as 1 in our experiments. The learning can be summarized as the following optimization problem:

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

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

We design the model by different kinds of loss functions. The standard supervised pixelwise cross entropy loss of Eq. (1) is chosen to train the generator on the supervised source domain data. While the adversarial loss of Eq. (2) is used to train the discriminator to classify the features coming from source or target domains. Eq. (3) is introduced to let the generator output domain-invariant features so as to confuse the discriminator. If the discriminator could not distinguish the features coming from the source or target domains, the generator outputs the domain-invariant features. During the training procedure, we alternately optimize the network parameters θ_G for $G(\cdot)$ by optimizing Eq. (4) and the network parameters θ_D for $D(\cdot)$ by optimizing Eq. (5). We see the loss decreasing in the network training, so the domain-invariant features are obtained by the generator.

D. 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 RGBN-SOD dataset into training, validation and testing subsets as the ratio of 5:1:4 same as the split in the MSRA-B dataset [81]. The image pairs are randomly selected from the simple and 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.

V. Experiment

A. Source-Domain Dataset and the Pseudo-NIR Images

For the unsupervised domain adaptation SOD task, we choose the MSRA-B dataset [76], [81] as our source domain, and our RGBN-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 by the ratio of 5:1:4 (training: 2500 images, validation: 500 images, testing: 2000 images) as that in [81]. For training the CycleGAN model, we take all the 2500 training images in MSRA-B and all the 780 NIR images in RGBN-SOD dataset as the two domains for the image transfer.

The generalization ability of different SOD datasets may be different [49]. The reason for us to choose MSRA-B dataset as the source domain dataset is that the average ground-truth of the salient objects in the MSRA-B dataset is similar to that of the proposed RGBN-SOD dataset, as shown in Fig. 3(c) and Fig. 3(d). From the average ground-truth maps, we see that 1) the salient object in these two datasets are mainly located around the center of the image; 2) the ratio of the salient object areas to the whole image are also similar in these two datasets. In the cases of extremely different area distributions in source and target domains such as too large, small, sparse object size and locations, the area distribution might cause damages to the deep neural networks model generalization. It is worth mentioning that the main domain discrepancy in this paper comes from the appearance feature differences between the source and target domains, e.g., color, texture, contrast, brightness, etc, not from the area distribution.

For training the CycleGAN model, we choose the cross-entropy loss mode, image buffer is set as 50 inspired by [77] 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. 5, and we can see that the generated pseudo-NIR images are reasonable and similar as the real NIR images.

B. Evaluation Metrics

We evaluate the proposed multi-spectral SOD method performance using Precision-Recall (PR) curve, maximum F-measure

(max-F), Mean Absolute Error (MAE), and Structural similarity measure (S-measure). We also evaluate the average precision, recall and F-measure with an adaptive threshold that is twice the mean value of the saliency map [6]. The value of F-measure is defined as $F_\gamma = \frac{(1+\gamma^2)\times Precision\times Recall}{\gamma^2\times Precision+Recall}$, where γ^2 is set to 0.3 as suggested in [51]. 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 θ. The PR curve of a dataset is computed from the average precision and recall value over the whole dataset. The MAE error [82] is calculated as the average absolute pixelwise difference between the predicted saliency map and the binary ground truth. The structural similarity measure (S-measure) [84] is a new metric proposed to evaluate the similarity between a non-binary saliency map and a ground-truth map. We calculate the S-measure as defined in [83], $S_\gamma = \gamma \times S_o + (1-\gamma) \times S_r$, where γ is set to 0.5 as suggested in [83], S_o represents the object-aware structural similarity and S_r is the region-aware structural similarity.

C. 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 pixel-level annotations to perform well on the RGBN-SOD dataset without pixel-level annotations. We implement our networks using PyTorch running on a single Tesla P40 GPU. The FCN8s network [78] using VGG16 is used as our backbone model, and we also conduct experiments on the FCN16 s 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 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 [84]. In our experiments, names "*8 s" or "*16 s" indicate the backbone network as FCN8s or FCN16 s 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 are 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 [85], 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 1×10^{-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" indicates 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 I and Table II.

Table I shows the SOD results compared with other methods. We compare our unsupervised method with some salient object detection methods as RC [8], LRK [44], CWS [45], FT [46], and DCL [51]. The first four methods are feature-based traditional methods and the last one is deep learning based method. Figure 8 shows sample results of different SOD methods and the PR curve of the related results are also shown in Fig. 9. From the results, we can find that the proposed method performs better than the other SOD methods on Precision, F-measure, MAE, maxF, and Smeasure metrics. Table II shows the performance change of each component of the proposed method. Taking FCN8s as

TABLE I
PERFORMANCE COMPARISONS TO OTHER METHODS ON THE RGBN-SOD DATASET. (FOR EACH COLUMN IN THE TABLE, THE TOP TWO BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND *BOLD ITALIC*, RESPECTIVELY.)

Method	Subset	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
RC [8]	RGB	0.6612	0.6812	0.6310	0.1621	0.7032	0.6930
	RGBN	0.6743	0.7785	0.6641	0.1455	0.7333	0.7242
LRK [44]	RGB	0.5865	0.6923	0.5640	0.1743	0.6588	0.6548
	RGBN	0.5892	0.7018	0.5658	0.1786	0.6640	0.6488
CWS [45]	RGB	0.5916	0.5625	0.5471	0.2428	0.6137	0.5991
	RGBN	0.5723	0.5225	0.5059	0.2452	0.5784	0.6079
FT [46]	RGB	0.3496	0.3954	0.3322	0.1974	0.3622	0.4926
	RGBN	0.3952	0.4320	0.3701	0.1934	0.4216	0.5252
DCL [51]	RGB	0.7789	0.7636	0.7461	0.0738	0.7885	0.7730
	RGBN	***0.7907***	**0.8436**	***0.7791***	0.0768	***0.8367***	0.7961
SOD16s$^+$	RGB^2	0.6946	0.7770	0.6806	0.0851	0.7502	0.7655
	RGBN	0.7209	***0.8259***	0.7137	0.0764	0.8093	***0.8011***
SOD8s$^+$	RGB^2	***0.7907***	0.7666	0.7572	***0.0692***	0.7966	0.7914
	RGBN	**0.8266**	0.8207	**0.8030**	**0.0611**	**0.8458**	**0.8281**

TABLE II
PERFORMANCE ON THE UNSUPERVISED DOMAIN ADAPTATION FOR THE MULTI-SPECTRAL SOD ON THE RGBN-SOD DATASET

Method	Subset	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
FCN16s	RGB	0.7096	0.7262	0.6773	0.0948	0.7106	0.7228
	RGBN	0.7201	0.7986	0.7079	0.0984	0.7632	0.7490
SOD16s	RGB^2	0.6007	0.7700	0.5998	0.1135	0.7032	0.7241
	RGBN	0.6380	**0.8367**	0.6444	0.0996	0.7773	0.7660
SOD16s$^+$	RGB^2	0.6946	0.7770	0.6806	0.0851	0.7502	0.7655
	RGBN	0.7209	***0.8259***	0.7137	0.0764	0.8093	0.8011
FCN8s	RGB	0.7737	0.7380	0.7321	0.0801	0.7650	0.7592
	RGBN	0.7875	0.8175	0.7695	0.0829	0.8194	0.7802
SOD8s	RGB^2	0.7613	0.7575	0.7320	0.0766	0.7688	0.7706
	RGBN	***0.8170***	0.8117	***0.7911***	0.0700	***0.8200***	***0.8022***
SOD8s$^+$	RGB^2	0.7907	0.7666	0.7572	***0.0692***	0.7966	0.7914
	RGBN	**0.8266**	0.8207	**0.8030**	**0.0611**	**0.8458**	**0.8281**

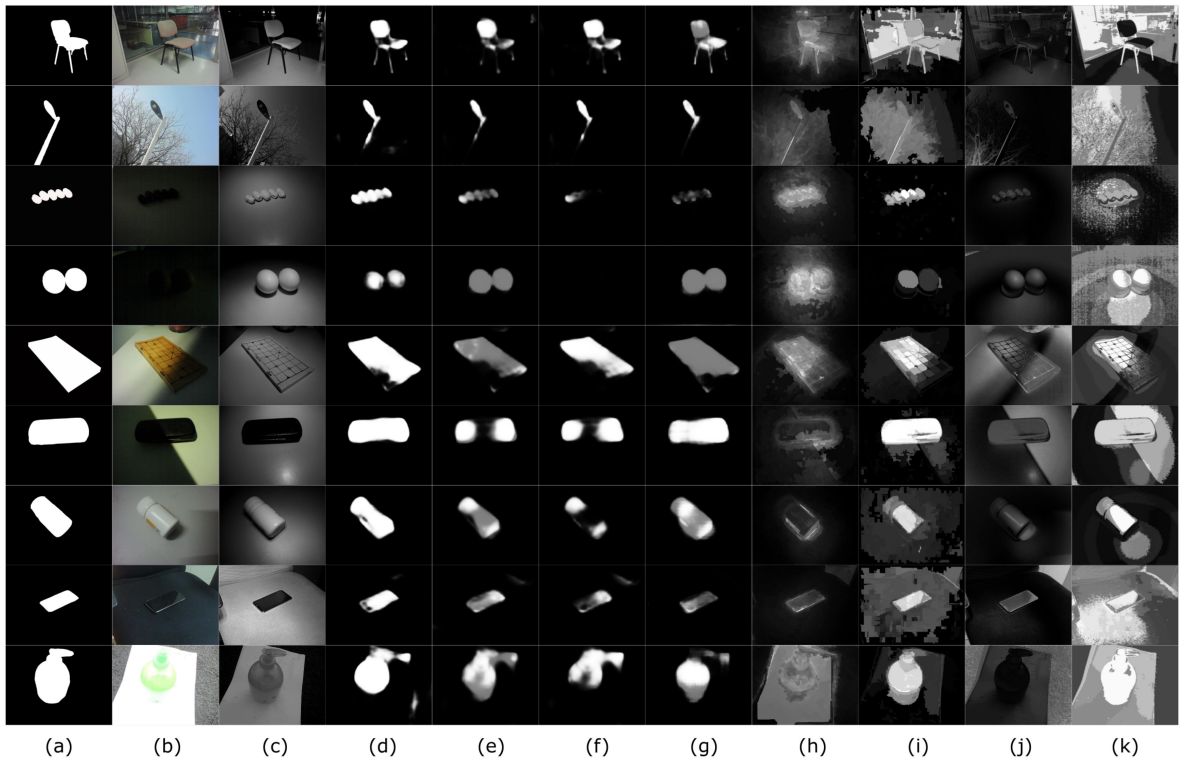


Fig. 8. Sample results of unsupervised multi-spectral salient object detection on the RGBN-SOD dataset: (a) ground-truth, (b) RGB image, (c) NIR image, (d) RGBN results on SOD8s$^+$ (the proposed method with unsupervised domain adaptation), (e-f) RGBN and RGB results on FCN8s, (g) DCL, (h) LRK, (i) RC, (j) FT, and (k) CWS. .

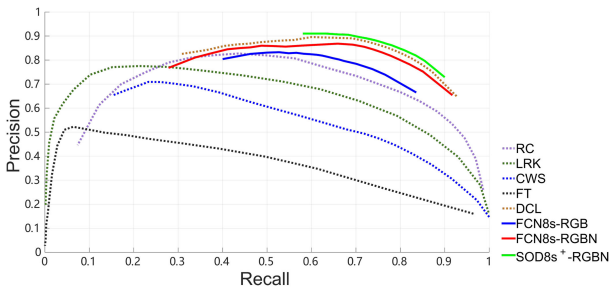


Fig. 9. PR curves of different unsupervised SOD methods on the proposed RGBN-SOD dataset.

TABLE III
PERFORMANCE ON THE SUPERVISED DOMAIN ADAPTATION FOR THE MULTI-SPECTRAL SOD ON THE RGBN-SOD DATASET

Method	Subset	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
sVGG16s	RGB^2	0.6976	0.8413	0.7016	0.0799	0.7705	0.7812
	RGBN	0.7689	0.8783	0.7728	0.0653	0.8303	0.8292
sSOD16s	RGB^2	0.7101	0.8654	0.7196	0.0716	0.8024	0.8154
	RGBN	0.7742	0.8970	0.7829	0.0570	0.8618	0.8593
sSOD16s$^+$	RGB^2	0.6807	0.8855	0.6991	0.0714	0.8075	0.8188
	RGBN	0.7385	0.9137	0.7559	0.0579	0.8661	0.8618
sVGG8s	RGB^2	0.7466	0.9083	0.7630	0.0586	0.8309	0.8422
	RGBN	0.7945	**0.9270**	0.8089	0.0464	0.8793	0.8812
sSOD8s	RGB^2	0.7588	0.9022	0.7702	0.0548	0.8433	0.8492
	RGBN	***0.8117***	***0.9238***	***0.8276***	***0.0421***	***0.8904***	***0.8878***
sSOD8s$^+$	RGB^2	0.7955	0.8904	0.8010	0.0498	0.8543	0.8586
	RGBN	**0.8502**	0.9156	**0.8533**	**0.0389**	**0.9031**	**0.8940**

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 FCN16 s as a baseline. In addition, both Table I and Table II 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.

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

Table III 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, the model initialized by the pre-trained model of "SOD8s$^+$" gets the best performance, i.e., maxF=0.9031, Smeasure=0.8940,

TABLE IV
CO-SALIENCY PERFORMANCE ON THE COLLECTED CO-RGBN DATASET USING DIFFERENT INITIALIZATIONS TO START THE HSCS METHOD [18]

Method	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
HSCS_DCMC	***0.6897***	0.8535	***0.6981***	***0.0694***	***0.7988***	***0.7899***
HSCS_RC	0.6542	***0.8764***	0.6721	0.0810	0.7945	0.7301
HSCS_SOD8s$^+$	**0.7752**	**0.9320**	**0.7903**	**0.0396**	**0.8989**	**0.8778**

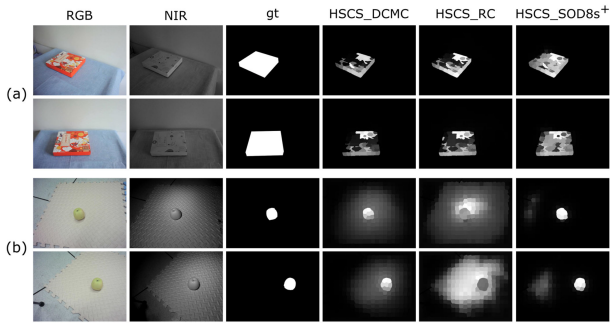


Fig. 10. Sample co-saliency detection results by three methods on the collected co-RGBN dataset.

TABLE V
PERFORMANCE ON OPTIMAL-SOD [42] DATASET CONTAINING 40 RGB-NIR IMAGE PAIRS

Method	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
RC [8]	0.3678	0.6914	0.3846	0.1899	0.4471	0.5807
sVGG8s	0.4892	***0.8056***	0.5132	***0.0765***	0.6120	**0.7571**
sSOD8s	***0.5147***	0.7898	***0.5341***	**0.0718**	***0.6182***	0.7242
sSOD8s$^+$	**0.5340**	**0.8700**	**0.5562**	0.0780	**0.6243**	***0.7425***

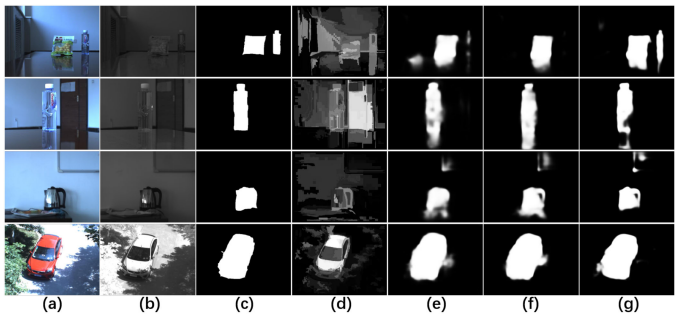


Fig. 11. Sample results of OPTIMAL-SOD dataset [42] with RGB-NIR image pairs: (a) RGB image, (b) NIR image, (c) ground-truth, (d) RC result, (e) sVGG8s result, (f) sSOD8s result, and (g) sSOD8s$^+$ result.

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.

E. Co-Saliency Performance on the co-RGBN Dataset

Co-saliency detection is to highlight the common salient object(s) in a group of images. Since there are not many researches on multi-spectral (RGBN) co-salient object detection, we simply explore the RGBN co-salient object detection with the collected co-RGBN dataset. In some methods designed for co-salient object detection like [18], [87], intra-saliency maps are needed for the initialization. In this section, we try to study the effects of the quality of the initialized intra-saliency map to start the RGBN co-saliency detection. We take the HSCS method [18] originally proposed for the RGBD co-saliency detection as the baseline framework for the RGBN co-salient object detection. The HSCS method needs initialized intra-saliency map for each individual image to start its co-saliency detection. We select the initialized intra-saliency maps by three different methods–DCMC [63], RC [8] and SOD8s $^+$–to start the HSCS method on the collected co-RGBN dataset, which are denoted as "HSCS_ DCMC," "HSCS _ RC," and "HSCS _ SOD8s $^+$," respectively. The co-saliency performance on the collected co-RGBN dataset using different methods is shown in Table IV. We can see that the proposed "HSCS_ SOD8s$^+$" provides useful cues for finding the co-saliency object(s) in a group of images and achieves the best performance. Figure 10 shows the sample results by the three methods.

F. Performance on the Existing Multi-Spectral SOD Dataset

As far as we know, the current largest public multi-spectral RGB-NIR SOD dataset (except our proposed RGBN-SOD dataset) is the OPTIMAL-SOD dataset collected by [42] with only 40 RGB-NIR image pairs in it. The OPTIMAL-SOD dataset contains indoor and outdoor images with the resolution of 512×384 pixels. One or several salient objects are included in each RGB-NIR image pairs.

Since the number of image pairs in the dataset provided by [42] is not sufficient to train a deep learning model, we directly take the whole OPTIMAL-SOD dataset as the testing set. The models trained on the training set of our RGBN-SOD dataset are directly adopted as the salient object detectors. The experimental results are shown in Table V. In Table V, the first line shows the result of pixelwise average combination of the RGB and NIR prediction results by the RC method [8], and the bottom three lines show the prediction results of the supervised models discussed in Section V-D with both RGB and NIR modalities as the input. The reason we choose the RC method as the comparison is that the RC method shows almost the best performance reported in [42]. It is obvious that the trained models under the supervision of RGBN-SOD dataset obtain better performances comparing to the RC method. The model initialized by the "SOD8s$^+$" still perform the best on the OPTIMAL-SOD dataset. Figure 11 shows the sample detection results on the OPTIMAL-SOD dataset. The "sSOD8s$^+$" model can get more accurate SOD detection results in Fig. 11. The RC method is a traditional contrast based salient object detection method. As mentioned in [42], images in OPTIMAL-SOD dataset has complex background by containing several disturbing objects of different colors and the salient objects in many images do not have a unique distinguishable color. Therefore, the contrast between the object and background in some images in this dataset is low. In this way, the contrast-based method RC may not be able to capture sufficient cues to detect the salient objects, leading to a low performance. However, the proposed method using deep learning is more robust by learning and using deeper features.

TABLE VI
PERFORMANCE OF THE MULTI-SPECTRAL SOD ON THE RGBN-SOD DATASET COMPARED TO STATE-OF-THE-ART SOD MODELS

Unsupervised Method	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
DTM_FCN8s [88]	0.7268	0.7784	0.7070	0.1150	0.7868	0.7522
BBS [90]	0.8746	0.9268	0.8759	***0.0352***	0.9195	0.9080
PGAR [89]	0.8592	0.9148	0.8589	0.0366	0.9015	0.8920
ASNet [59]	0.8227	0.9459	0.8369	0.0416	0.9058	0.9006
BAS [53]	0.8347	0.9298	0.8418	0.0415	0.8747	0.8938
CSF+Res2Net [61]	0.8556	0.9462	0.8643	0.0378	0.9043	**0.9192**
DSS [87]	0.8350	0.7930	0.7947	0.0652	0.8415	0.7994
EGNet [57]	0.8371	**0.9537**	0.8498	0.0375	0.9066	0.9118
PAGE [58]	0.8345	***0.9478***	0.8481	0.0373	0.8958	0.9072
CPD [52]	0.8392	0.9110	0.8392	0.0478	0.8827	0.8768
CPD_SOD	0.8673	0.8821	0.8538	0.0414	0.8868	0.8746
CPD_SOD$^+$	0.8624	0.9034	0.8569	0.0395	0.8920	0.8735
CPD_ResNet [52]	0.8090	0.9438	0.8219	0.0454	0.9087	0.8988
CPD_ResNet_SOD	**0.9035**	0.8974	**0.8907**	***0.0352***	***0.9229***	0.9006
CPD_ResNet_SOD$^+$	***0.8936***	0.9174	***0.8887***	**0.0316**	**0.9249**	***0.9122***
Supervised Method	**Precision**	**Recall**	**Fmeasure**	**MAE**	**maxF**	**Smeasure**
sDTM_SOD8s$^+$ [88]	0.7332	0.8203	0.7202	0.1109	0.8108	0.7749
sBBS [90]	0.8857	**0.9557**	0.8923	0.0373	0.9375	0.9194
sBAS [53]	0.8065	0.9340	0.8190	0.0476	0.8953	0.8921
sCSF+Res2Net [61]	0.6605	0.8915	0.6785	0.0905	0.8115	0.8106
sDSS [87]	0.8470	0.8973	0.8464	0.0531	0.8834	0.8550
sEGNet [57]	0.8303	***0.9493***	0.8414	0.0403	0.9086	0.9105
sCPD [52]	0.8790	0.9060	0.8719	0.0389	0.9079	0.8945
sCPD_SOD	0.8704	0.9390	0.8745	0.0330	0.9213	0.9113
sCPD_SOD$^+$	0.8868	0.9438	0.8891	0.0303	0.9261	0.9175
sCPD_ResNet [52]	0.8927	0.9447	0.8979	***0.0267***	0.9276	0.9246
sCPD_ResNet_SOD	**0.9083**	0.9447	**0.9078**	***0.0267***	***0.9386***	***0.9279***
sCPD_ResNet_SOD$^+$	***0.9033***	0.9489	***0.9065***	**0.0254**	**0.9393**	**0.9306**

TABLE VII
PERFORMANCE ON UNSUPERVISED DOMAIN ADAPTATION FOR THE MULTI-SPECTRAL SOD USING FCN8S MODEL ON THE RGBN-SOD DATASET WITH DIFFERENT SCALES OF SOURCE DOMAIN IMAGE NUMBER

Method	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
SOD8s$_{2500}$	0.8170	0.8117	0.7911	0.0700	0.8200	0.8022
SOD8s$^+_{2500}$	***0.8266***	0.8207	***0.8030***	***0.0611***	***0.8458***	0.8281
SOD8s$_{4500}$	0.8118	**0.8448**	0.7979	0.0618	0.8369	***0.8302***
SOD8s$^+_{4500}$	**0.8275**	***0.8304***	**0.8057**	**0.0596**	**0.8473**	**0.8323**

G. Discussion

1) Applying the Proposed Method to the State-of-The-Art SOD Model: We take the state-of-the-art SOD model, Cascaded Partial Decoder (CPD) framework proposed in [52], as the backbone model to test the proposed deep domain adaptation method.

Firstly, we choose CPD with VGG16 model as its backbone. We fine-tune the proposed method under the unsupervised situation with CPD model as the backbone. During fine-tuning, the default parameter setting of the CPD model is adopted. "CPD" indicates fine-tuning the original CPD model with the MSRA-B dataset as the training set in the same way with the FCN method in our experiment. We extend the CPD model as a two-branch network in the same way as we extend FCN network mentioned in Section IV-B, and we indicate the two-branch CPD model as "CPD _ SOD". We also apply the proposed adversarial domain adaptation framework on "CPD _ SOD" and name it as "CPD _ SOD$^+$". Then, we also choose CPD with ResNet50 model as its backbone. The experiment settings are the same as the VGG16 model based CPD method above, and the original CPD with Resnet50 model as its backbone is indicated as "CPD _ ResNet," the two-branch CPD model is indicated as "CPD _ ResNet_ SOD" and the model applied the proposed adversarial domain adaptation framework on "CPD _ ResNet_ SOD" is indicated as "CPD _ ResNet_ SOD$^+$". Table VI shows the results by taking RGB and NIR image pairs as the input.

From Table VI, we can see that the proposed method using CPD model as the backbone shows better performance than the proposed method with FCN8s model as the backbone. On one hand, it shows that the proposed methods performance could be improved by using a more advanced backbone model, e.g., the CPD model. On the other hand, the proposed domain adaptation method could also improve the maxF performance on CPD model from 0.8827 to 0.8920, and on CPD _ ResNet model from 0.9087 to 0.9249. We also add supervised experiments on the CPD backbone with VGG16 and ResNet50 model, receptively. With CPD as the backbone, supervised models can obtain better performance than the unsupervised models.

Besides the CPD method, we also compare with other deep learning based SOD methods, such as BAS [53], ASNet [59], [60], DSS [88], PAGE [58], CSF+Res2Net [61], [62], EGNet [57] and RGBD SOD methods, such as DTM [88], PGAR [89], BBS [89]. From Table VI, we can find that most of the deep learning based SOD methods can get better performance than the DTM method, but their performance is lower than the proposed method with CPD as the backbone in both unsupervised and supervised settings on most of the evaluation metrics. We can also see that directly applying the RGBD SOD method DTM to multi-spectral SOD could not obtain satisfactory performance.

2) Influence on the Scale of Source Domain Image Number: We also add an experiment to study the influence of the image number of the training set under the unsupervised domain adaptation setting. The default setting is using 2500 training images in MSRA-B as the source domain in all the above experiments. We enlarge the number of images in the source domain by taking all the training and testing images in MSRA-B (totally 4500 images) as the source domain for training. Table VII shows the RGB-NIR image pair's performance trained with different scales of source domain image number. Compared to 2500 images as the source domain, when taking the FCN8s as our baseline model, the maxF is improved from 0.8200 using 2500 images to 0.8369 using 4500 images. After the proposed domain adaptation, the maxF is slightly improved from 0.8458 using 2500 images to 0.8473 using 4500 images. We can see that the domain adaptation performance is slightly improved with more source domain images. It also shows that the proposed domain adaptation method is robust no matter using 2500 images or 4500 images as the source domain.

3) Evaluation on the Enlarged RGBN-3000 Dataset: In this section, we enlarge the dataset to evaluate the performance of the proposed method. First, we enlarge the dataset into 1000 image pairs by adding 220 newly collected RGBN image pairs to the original 780 RGBN image pairs. Then we flip all the images in horizontal direction to generate additional 1000 RGBN image pairs, and add "salt and pepper" noise with 0.2 noise density to generate another 1000 RGBN image pairs. Combining them together, we get a larger-scale dataset, 3000 RGBN image pairs, which is named as "RGBN-3000" dataset in this paper. For the unsupervised experiment setting, all the 3000 image pairs in the RGBN-3000 dataset are taken as the target domain. For the

TABLE VIII
PERFORMANCE OF THE MULTI-SPECTRAL SOD ON THE RGBN-3000 DATASET WITH CPD [52] AS BASELINE

Unsupervised Method	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
CPD [52]	0.8126	***0.8566***	0.7997	0.0583	0.8482	0.8243
CPD_SOD	**0.8545**	0.8450	***0.8290***	***0.0458***	***0.8613***	***0.8452***
CPD_SOD^{+}	***0.8434***	**0.8712**	**0.8299**	**0.0435**	**0.8665**	**0.8567**
Supervised Method	Precision	Recall	Fmeasure	MAE	maxF	Smeasure
sCPD [52]	***0.8844***	0.9458	***0.8891***	0.0277	0.9304	0.9229
sCPD_SOD	0.8815	**0.9536**	0.8888	**0.0263**	***0.9350***	***0.9281***
sCPD_SOD^{+}	**0.8860**	***0.9490***	**0.8912**	***0.0265***	**0.9353**	**0.9297**

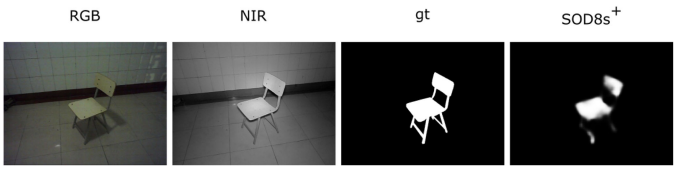


Fig. 12. Sample failure case of the proposed SOD8s^{+} method on the RGBN-SOD dataset.

supervised setting, we split the dataset in ratio 5:1:4 by following the rule mentioned above.

Table VIII shows the results on the RGBN-3000 dataset. In this part, we adopt the same experiment settings as in Section V-G1. "CPD" indicates fine-tuning the original CPD model with the MSRA-B training set, "CPD _ SOD" indicates the extended two-branch CPD model, and "CPD _ SOD^{+}" indicates the model applied with the proposed adversarial domain adaptation framework on "CPD _ SOD" model. The results in Table VIII show that the proposed method can improve the performance on maxF and Smeasure evaluations under both unsupervised and supervised settings.

4) Failure Case: We show a failure case of the proposed method in Fig. 12. In this example, the main parts of the salient object (i.e., chair) have been discovered by the proposed method. However, the proposed method might ignore the tiny-thin objects like the chair legs as shown in Fig. 12.

VI. CONCLUSION

In this paper, we systematically studied the multi-spectral salient object detection problem. We first proposed a new large dataset RGBN-SOD 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 in deep learning 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 experimental results and discussions show that the proposed deep domain adaptation method is also helpful to improve the saliency detection accuracy. In addition, we also collected a new dataset co-RGBN including 200 RGB-NIR image pairs to study the multi-spectral co-salient object detection problem in this paper. Future work will be focused on continually improving the performance on the collected multi-spectral datasets.

ACKNOWLEDGMENT

The authors gratefully appreciate the help of Dingxin Yan for image capturing.

REFERENCES

[1] S. Song *et al.*, "Multi-spectral salient object detection by adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12023–12030.
[2] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.
[3] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1354–1362.
[4] F. Sun and W. Li, "Saliency guided deep network for weakly-supervised image segmentation," *Pattern Recognit. Lett.*, vol. 120, pp. 62–68, 2019.
[5] P. Zhang *et al.*, "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *Pattern Recognit.*, vol. 100, 2020, Art. no. 107130.
[6] H. Yu, K. Zheng, J. Fang, H. Guo, W. Feng, and S. Wang, "Co-saliency detection within a single image," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7509–7516.
[7] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1896–1909, Sep. 2016.
[8] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.
[9] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.
[10] Y. Tang and X. Wu, "Salient object detection using cascaded convolutional neural networks and adversarial learning," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2237–2247, Sep. 2019.
[11] S. Song *et al.*, "An easy-to-hard learning strategy for within-image co-saliency detection," *Neurocomputing*, vol. 358, pp. 166–176, 2019.
[12] H. Yu, K. Zheng, J. Fang, H. Guo, and S. Wang, "A new method and benchmark for detecting co-saliency within a single image," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3051–3063, Dec. 2020.
[13] S. Liu and Z. Liu, "Multi-channel cnn-based object detection for enhanced situation awareness," 2017, *arXiv:1712.00075*.
[14] D. K. Shin, M. U. Ahmed, and P. K. Rhee, "Incremental deep learning for robust object detection in unknown cluttered environments," *IEEE Access*, vol. 6, pp. 61 748–61760, 2018.
[15] A. Singha and M. K. Bhowmik, "TU-VDN: Tripura university video dataset at night time in degraded atmospheric outdoor conditions for moving object detection," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2936–2940.
[16] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.
[17] J. Tang, L. Jin, Z. Li, and S. Gao, "RGB-D object recognition via incorporating latent data structure and prior knowledge," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1899–1908, Nov. 2015.
[18] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, "HSCS: Hierarchical sparsity based co-saliency detection for rgbd images," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1660–1671, Jul. 2019.
[19] S. Matsui, T. Okabe, M. Shimano, and Y. Sato, "Image enhancement of low-light scenes with near-infrared flash images," *Inf. Media Technol.*, vol. 6, no. 1, pp. 202–210, 2011.
[20] X. Shen, Q. Yan, L. Xu, L. Ma, and J. Jia, "Multispectral joint image restoration via optimizing a scale map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2518–2530, Dec. 2015.
[21] C.-H. Son and X.-P. Zhang, "Near-infrared image dehazing via color regularization," 2016, *arXiv:1610.00175*.
[22] L. Schaul, C. Fredembach, and S. Ssstrunk, "Color image dehazing using the near-infrared," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 1629–1632.

[23] X. Wang, F. Dai, Y. Ma, J. Guo, Q. Zhao, and Y. Zhang, "Near-infrared image guided neural networks for color image denoising," in *IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3807–3811.

[24] M. Brown and S. Ssstrunk, "Multi-spectral sift for scene category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 177–184.

[25] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019.

[26] J. Jiang, X. Feng, F. Liu, Y. Xu, and H. Huang, "Multi-spectral RGB-NIR image classification using double-channel cnn," *IEEE Access*, vol. 7, pp. 20 607–20613, 2019.

[27] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 627–639, Apr. 2007.

[28] A. V. Vanmali and V. M. Gadre, "Visible and NIR image fusion using weight-map-guided Laplacian-Gaussian pyramid for improving scene visibility," *S*vol. 42, no. 7, pp. 1063–1082, 2017.

[29] S. Liu and G. Tian, "An indoor scene classification method for service robot based on CNN feature," *J. Robot.*, vol. 2019, 2019.

[30] H.-Z. Chen, G.-H. Tian, and G.-L. Liu, "A selective attention guided initiative semantic cognition algorithm for service robot," *Int. J. Automat. Comput.*, vol. 15, no. 5, pp. 559–569, 2018.

[31] L. Du *et al.*, "Adaptive visual interaction based multi-target future state prediction for autonomous driving vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4249–4261, May 2019.

[32] S. Shashikar and V. Upadhyaya, "Traffic surveillance and anomaly detection using image processing," in *Proc. Int. Conf. Image Inf. Process.*, 2017, pp. 1–6.

[33] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[34] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 136–145.

[35] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5455–5463.

[36] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2960–2967.

[37] Y. Lin *et al.*, "Cross-domain recognition by identifying joint subspaces of source domain and target domain," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1090–1101, Apr. 2017.

[38] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, "Domain adaptation for convolutional neural networks based remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1324–1328, Aug. 2019.

[39] D. Guo, Y. Pei, K. Zheng, H. Yu, Y. Lu, and S. Wang, "Degraded image semantic segmentation with dense-gram networks," *IEEE Trans. Image Process.*, vol. 29, pp. 782–795, 2019.

[40] J.-A. Bolte *et al.*, "Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1404–1413.

[41] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[42] Q. Wang, G. Zhu, and Y. Yuan, "Multi-spectral dataset and its application in saliency detection," *Comput. Vis. Image Understanding*, vol. 117, no. 12, pp. 1748–1754, 2013.

[43] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 34–41, 2013.

[44] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 853–860.

[45] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[46] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

[47] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based image segmentation with connectivity priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[48] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 29–42.

[49] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," 2019, *arXiv:1904.09146*.

[50] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating Multi-Level Convolutional Features for Salient Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[51] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 478–487.

[52] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3907–3916.

[53] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.

[54] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked U-shape network with channel-wise attention for salient object detection," IEEE Trans. Multimedia, to be published.

[55] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. Lu, "A multistage refinement network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3534–3545, 2020.

[56] S. Zhou *et al.*, "Hierarchical U-shape attention network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8417–8428, 2020.

[57] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet:Edge guidance network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8778–8787.

[58] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.

[59] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1711–1720.

[60] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1913–1927, Aug. 2020.

[61] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100 k parameters," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 702–721.

[62] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[63] R. Cong, J. Lei, C. Zhang, Q. Huang, X. C ao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, Jun. 2016.

[64] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[65] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[66] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, Feb. 2018.

[67] J. Zhang *et al.*, "Uc-net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8582–8591.

[68] N. Salamati, C. Fredembach, and S. Ssstrunk, "Material classification using color and NIR images," in *Color Imag. Conf.*, vol. 2009, no. 1. *Soc. Imag. Sci. Technol.*, 2009, pp. 216–222.

[69] S. He and R. W. Lau, "Saliency detection with flash and no-flash image pairs," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 110–124.

[70] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2019, pp. 141–146.

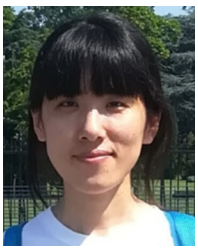
[71] X. Zhu, X. Xu, and N. Mu, "Saliency detection based on the combination of high-level knowledge and low-level cues in foggy images," *Entropy*, vol. 21, no. 4, p. 374, 2019.

[72] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[73] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.

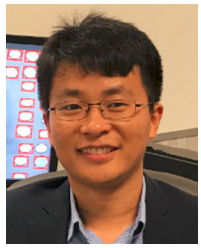
[74] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Prez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2517–2526.

[75] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, p. 1369, 2019.
[76] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, 2017.
[77] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[78] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[80] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[81] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2083–2090.
[82] F. Perazzi, P. Krhenbhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 733–740.
[83] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-Measure: A. New Way to Evaluate Foreground Maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.
[84] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[85] K. Wada, "PyTorch Implementation of Fully Convolutional Networks," 2017. [Online]. Available: https://github.com/wkentaro/pytorch-fcn
[86] H. Yu, M. Xian, and X. Qi, "Unsupervised co-segmentation based on a new global gmm constraint in mrf," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 4412–4416.
[87] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.
[88] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.
[89] S. Chen and Y. Fu, "Progressively guided alternate refinement network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020.
[90] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 275–292.



Shaoyue Song received the B.E. degree, in 2014 from Beijing Jiaotong University, Beijing, China, where she is currently working toward the Ph.D. degree. In 2018, she was a Visiting Student with the University of South Carolina, Columbia, SC, USA, supported by the China Scholarship Council. Her current research interests include salient object detection and image classification.

Zhenjiang Miao (Member, IEEE) received the B.E. degree from Tsinghua University, Beijing, China, in 1987, and the M.E. and Ph.D. degrees from Northern Jiaotong University, Beijing, China, in 1990 and 1994, respectively. From 1995 to 1998, he was a Postdoctoral Fellow with the école Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications, Institut National Polytechnique de Toulouse, Toulouse, France, and was a Researcher with the Institute National de la Recherche Agronomique, Sophia Antipolis, Biot, France. From 1998 to 2004, he was with the Institute of Information Technology, National Research Council Canada, Nortel Networks, Ottawa, ON, Canada. In 2004, he joined Beijing Jiaotong University, Beijing, China. He is currently a Professor, Director with Media Computing Center, Beijing Jiaotong University, and the Director with the Institute for Digital Culture Research, Center for Ethnic & Folk Literature & Art Development, Ministry Of Culture, China. His current research interests include image and video processing, multimedia processing, and intelligent human-machine interaction.



Hongkai Yu (Member IEEE) received the Ph.D. degree in computer science and engineering from University of South Carolina, Columbia, SC, USA, in 2018. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH, USA. His research interests include computer vision, machine learning, deep learning, and intelligent transportation system.

Jianwu Fang (Member IEEE) received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an Associate Professor, the Founder and Director with the Laboratory of Traffic Vision Safety, College of Transportation Engineering, Chang'an University, Xi'an, China. He has authored or coauthored many papers on top-ranked journals and conferences, such as IEEE the TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Association for the Advancement of Artificial Intelligence, International Conference on Robotics and Automation*. His research interests include computer vision and pattern recognition.

Kang Zheng received the B.E. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2012 and the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2019. He is currently a Senior Research Scientist with PAII Inc. His current research interests include computer vision, deep learning, and medical image analysis.

Cong Ma received the B.E. degree, in 2013 from Beijing Jiaotong University, Beijing, China, where he is currently working toward the Ph.D. degree. In 2018, he was a Visiting Student with the University of California Merced, Merced, CA, USA, and supported by the China Scholarship Council. His current research interests include visual tracking and video analysis.

Song Wang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2002. From 1998 to 2002, he was a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, where he is currently a Professor. His current research interests include computer vision, image processing, and machine learning. He is currently the Publicity or Web Portal Chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society and an Associate Editor for IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition Letters*, and *Electronics Letters*. He is a member of the IEEE Computer Society.