RGB-'D' Saliency Detection With Pseudo Depth

Xiaolin Xiao, Yicong Zhou, *Senior Member, IEEE*, and Yue-Jiao Gong, *Member, IEEE*

***Abstract*— Recent studies have shown the effectiveness of using depth information in salient object detection. However, the most commonly seen images so far are still RGB images that do not contain the depth data. Meanwhile, the human brain can extract the geometric model of a scene from an RGB-only image and hence provides a 3D perception of the scene. Inspired by this observation, we propose a new concept named RGB-'D' saliency detection, which derives pseudo depth from the RGB images and then performs 3D saliency detection. The pseudo depth can be utilized as image features, prior knowledge, an additional image channel, or independent depth-induced models to boost the performance of traditional RGB saliency models. As an illustration, we develop a new salient object detection algorithm that uses the pseudo depth to derive a depth-driven background prior and a depth contrast feature. Extensive experiments on several standard databases validate the promising performance of the proposed algorithm. In addition, we also adapt two supervised RGB saliency models to our RGB-'D' saliency framework for performance enhancement. The results further demonstrate the generalization ability of the proposed RGB-'D' saliency framework.**

***Index Terms*— RGB-'D' saliency, pseudo depth, salient object detection.**

I. Introduction

SALIENT object detection aims at identifying the most distinctive and informative regions that grab human attention in images or videos [1]. Recently, this area has witnessed intensive studies and extensive applications to a variety of image processing and computer vision tasks, such as object recognition [2], image segmentation [3], visual tracking [4], and image and video compression [5].

Existing methods can be categorized into unsupervised and supervised approaches. Unsupervised methods [6]–[18] identify the salient objects through low-level feature extraction from the intrinsic cues of a specific image. They are usually computationally efficiency. Supervised learning methods [19]–[27] always require sufficient labeled images for the training procedure. These algorithms generally perform better than the unsupervised models. However, it is tough to manually label the salient objects in images and is usually time-consuming to train the models.

With the emergence of new cameras and laser scanners such as *Microsoft Kinect* and *SICK LMS 291*, RGB-D salient object detection has received increasing attention [28]–[31]. Intuitively, it would be very useful to utilize the depth information of images to identify the salient objects, since depth provides rich information on scene layout, shapes of objects, and other 3D cues [29], [30]. This information is highly consistent with human perception that can help to discriminate the foreground objects from the background. However, currently there are some limitations for 3D saliency detection. For example, the collection of RGB-D saliency databases is more expensive than that of RGB saliency databases, especially in outdoor scenes. This is because the widely used *Microsoft Kinect* is more suitable to capture indoor scenes due to the restriction of depth of field. When it comes to outdoor scenes, more expensive laser scanners are needed, e.g., *SICK LMS 291*. In addition to the challenge of data collection, the processing of the physical depth is complicated. Due to the different positions of the depth sensors and the color cameras, the raw depth and color images are in different coordinate systems and should be projected into the same coordinate space for alignment [32]. More importantly, there always exist missing or erroneous holes and regional gaps on the raw depth images. Such types of errors may come from the random noise of sensors, object reflection, shadows of the light patterns, etc. To avoid the propagation of errors into subsequent processing, the inpainting of the raw depth maps is necessary and the task itself is challenging [33]. Overall, these problems restrict the usage of the RGB-D images, and hence RGB saliency still dominates practical applications.

Considering the above issues, this paper proposes a new concept that derives a pseudo depth cue to assist saliency detection. The proposed pseudo depth measure is proper for both indoor and outdoor scenes. The adjective 'pseudo' here indicates that the depth is not the practical data sensed by physical devices, but instead it is estimated from the RGB images and is consistent with human perception. We name algorithms based on this concept as *RGB-'D' Saliency Detection*, where many opportunities of developing new and effective saliency models are opened up under this RGB-'D' saliency framework. Specifically, we develop a pseudo depth measure, named *semi-inverse image depth*, which is robust for estimating the scene depth in various types of images. This measure provides unique benefits in saliency detection since,

Manuscript received June 6, 2018; revised October 1, 2018; accepted November 2, 2018. Date of publication November 19, 2018; date of current version January 16, 2019. This work was supported in part by the National Natural Science Foundation of China under Grants 61873095 and U1701267, in part by the Macau Science and Technology Development Fund under Grant FDCT/189/2017/A3, and in part by the Research Committee at University of Macau under Grants MYRG2016-00123-FST and MYRG2018-00136-FST. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Khan M. Iftekharuddin. *(Corresponding author: Yue-Jiao Gong.)*

X. Xiao and Y. Zhou are with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: shellyxiaolin@gmail.com; yicongzhou@um.edu.mo).

Y.-J. Gong is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, South China University of Technology, Guangzhou 510006, China (e-mail: gongyuejiao@gmail.com).

Digital Object Identifier 10.1109/TIP.2018.2882156

in most cases, the salient objects capturing human attention are foreground objects in the images.

As an illustration, we develop a new RGB-'D' salient object detection algorithm exploiting the proposed pseudo depth. We will show that, compared with the widely adopted color and texture cues, the pseudo depth is informative and reliable, especially for the challenging images that endure low contrast, complex background and structures. In the proposed algorithm, a new background prior based on the pseudo depth is proposed. We also devise a foreground contrast map via aggregating the color, texture, and pseudo depth features. The foreground contrast and background prior are then fused together, through minimizing the data and smoothness loss of a cost function. We keep the algorithm work in a simple, intuitive, and straightforward way to observe the effectiveness of the RGB-'D' saliency detection. Experiments on several benchmark databases verify that the proposed algorithm achieves the state-of-the-art performance. In addition to designing the above algorithm, we also incorporate the RGB-'D' saliency detection concept into two existing RGB saliency algorithms. The performance enhancement observed in both algorithms validates the good generalization ability of the proposed RGB-'D' saliency framework.

Generally summarized, the novel contributions of this paper lie in the following aspects:

1) We propose the new concept of *RGB-'D' Saliency Detection*. The letter 'D' is enclosed in the quotation marks since it does not indicate the real depth data of an RGB-D image. Instead, we first estimate the depth information from an RGB image, and then apply the pseudo depth as a complementary cue to boost the performance of RGB saliency models. This new concept opens up opportunities of developing new and powerful saliency detection algorithms, either in a supervised or unsupervised way.

2) We develop a pseudo depth measure named *Semi-inverse Image Depth*. This pseudo depth measure successfully captures the scene depth in various types of images and provides a perceptually consistent cue to discriminate the salient objects from the background.

3) Based on the pseudo depth measure, we devise a new saliency prior named *'Pseudo-Depth-Driven Background Prior'*. The prior provides a robust characterization of the background probability of the pixels in an image.

4) As an illustration of the proposed RGB-'D' saliency framework, an algorithm termed *'Saliency Detection based on Pseudo Depth Prior (PDP)'* is proposed. The algorithm fuses the proposed background prior and a foreground contrast map together to obtain the saliency map. It achieves promising experimental results compared with the state-of-the-arts. The success of PDP verifies the advantages of using the proposed RGB-'D' saliency concept for developing powerful saliency detection algorithms with hand-crafted features.

5) We also adapt two supervised RGB saliency models to the RGB-'D' saliency framework. The experimental results further validate the effectiveness and the generalization ability of the proposed framework.

In the rest of this paper, Section II reviews the background knowledge. Section III poses a simple yet effective pseudo depth measure. Based on that, in Section IV, we define the concept of RGB-'D' saliency detection and propose a new algorithm for salient object detection based on this concept. In Section V, we compare the proposed algorithm with the state-of-the-arts. Then we exploit the RGB-'D' framework to improve the performance of existing RGB saliency models in Section VI. Finally, conclusions are drawn in Section VII.

II. Related Work

In this section, we briefly review the literature on salient object detection and depth estimation methods.

A. Salient Object Detection

Salient object detection is a popular research area that many approaches have been developed in the last decade. Generally, existing models can be categorized into two groups: RGB saliency and RGB-D saliency.

1) RGB Models With Intrinsic Cues: The unsupervised models identify the salient objects by considering the intrinsic cues of each input image only.

The pixel/regional **contrast** has been widely used since it is highly consistent with the human perception system in identifying the most distinctive object(s) in an image. Saliency Filters (SF) was designed in [7] by exploiting the color and spatial contrast of small regions. Cheng *et al*. [12] propose a more accurate evaluation of global contrast by defining the histogram based Regional Contrast (RC). Although it is perceptually intuitive, the contrast cue may suffer problems when the salient object(s) and the background are similar in appearance [6].

To address the limitation of the contrast cue, many algorithms concentrate on exploring effective **priors** and have shown promising results. The background priors are used to correctly reject the non-salient part of an image. For example, the boundary prior [6], [9], [10], [13] and the boundary connectivity prior were developed (e.g., Geodesic Saliency (GS) [6] and Robust Background Detection (RBD) [34]). In addition, the center prior [8], focusness and objectiveness priors [11] have been proposed. Usually, these priors become less effective when the specific assumptions are violated. For instance, when the salient objects touch the image borders, the performance of the boundary prior decreases [35].

Another popular category exploits the **diffusion-based** techniques by propagating the saliency information in an image graph. Algorithms in this class can generate visually smooth saliency maps due to the diffusion scheme. A graph-based Manifold Ranking (MR) [9] method was proposed to rank the saliency scores of both foreground and background seeds and these scores were fused to get the final saliency map. Jiang *et al*. [10] use the absorbing time of a Markov Chain (MC) to evaluate the saliency values of graph nodes. Later, new propagation schemes were developed using Cellular Automata (CA) [15], [16] and Minimum Spanning Tree (MST) [17], respectively. The diffusion process may also incorrectly suppress the salient region when the salient object touches image borders. Recently, Lu *et al*. [13]

use the pixel-level dense and sparse reconstruction (DSR) errors and Bayesian integration to overwhelm this limitation, and Zhou *et al*. [18] solve this problem via Diffusion on a Sparse Graph (DSG). These methods may fail to identify the whole salient object(s) since most of the graph-based methods ignore the consistency among different parts of the salient objects [11], [36].

Further, **hierarchical-segmentation-based** saliency detection was proposed in [37] (HS) and now has been commonly adopted for improving the detection robustness to salient objects with different scales. The studies in [13], [14], [16], and [36]–[38] exploit multi-scale segmentation and then fuse the saliency maps from different scales. Meanwhile, priors from different scales (e.g., object prior, focusness prior, spatial distribution prior) have been integrated with existing cues to assist saliency detection [8], [11], [36]. The hierarchy-based models show performance enhancement over their single-scale counterparts, and the idea of integrating the results from multiple scales can be easily extended to any single-scale models.

2) RGB Models With Extrinsic Cues: Some recently proposed methods utilize the extrinsic cues, either in the way of supervised learning or searching similar images [39]. For supervised learning, the studies in [19], [20], and [40] focus on feature integration. They learn the saliency scores from hand-crafted features. Later, the works in [21]–[24], [41], and [42] integrate the convolutional neural network features into this framework, in order to learn deep feature representation and feature integration simultaneously. In contrast, many other methods exploit different learning strategies to fuse the weak saliency models. A pioneering work [19] learns to combine different features using conditional random field. Then, Tong *et al*. [25] and Lu *et al.* [26] design a Bootstrap Learning (BL) algorithm to combine weak saliency models. Huang *et al*. [27] adopt the object proposal algorithm to generate candidate instances, and then find a decision boundary via Multiple Instances Learning (MIL). Generally, models using extrinsic cues have better performance compared with those using intrinsic cues, while being computationally more expensive. Besides, the extrinsic-cue-based models rely heavily on the manually labeled ground truth.

3) RGB-D Models: In addition to the RGB saliency models, depth information has shown a beneficial effect on saliency detection. Many studies have proposed RGB-D saliency models by incorporating the depth cue with RGB images. These models can also be classified into unsupervised algorithms [28]–[30] and supervised ones [31], [43]. Usually, the RGB-D saliency models outperform their RGB counterparts since the depth map provides rich information on scene layout, shapes of objects, and other 3D cues [29], [30]. This makes the RGB-D models feasible to pop out the salient objects even if they have similar appearances with the background. However, existing RGB-D models can only deal with physical depth, which has restricted practical applications.

B. The Pseudo Depth

Estimating depth map from RGB-only images has received considerable attention in computer vision. In terms of real scenarios, the human brain possesses a depth perception system integrating a variety of cues with respect to RGB-only images. Intuitively, the existence of edges, junctions and illumination variations in an RGB image may perceptively provide a 3D model of the scene [44]. Hence, we can extract the pseudo depth from RGB images in the absence of the real depth data to boost saliency detection models.

Generally, the pseudo depth of a scene can be estimated from multiple images or a single image. Schechner *et al*. [45] exploit a polarization-based method for haze removal as well as generating a pseudo depth map. It requires at least two input images taken with different degrees of polarization. He and Yuille [46] adopt a motion estimation algorithm to find the correspondence among multiple image frames and then generate the pseudo depth. Microsoft launched a platform AirSim [47] as an open source simulator for autonomous vehicles. When the real depth data of the input images are not available, AirSim can estimate the depth images from multiple images using stereo algorithms, e.g., [48]. Although these methods are effective in different scenarios, they are not applicable to our application since we focus on single image saliency detection.

Compared with multiple-image-based approaches [45], [46], single image depth estimation is more challenging since an RGB image may correspond to different scales of real scenarios [44]. To reliably estimate depth from a single image, existing works use feature consistency [49], multi-scale image features [50], image structures models [51], and deep neural networks [52], [53], etc. Among them, Zhang *et al*. [49] assume uniform color and texture for the scene and Delage *et al*. [51] exploit the geometry of floor and walls, which are suitable for indoor scenes and are not effective for outdoor images. Meanwhile, the model in [52] uses specific priors of outdoor scenes and is not suitable for indoor images. In addition, many other works are based on high-level features (e.g., the trained predictors) and are time-consuming in practice.

III. Pseudo Depth From Medium Transmission Model

As this paper aims at deriving a pseudo depth cue for salient object detection, we prefer a low-level depth model which is efficient to compute and is compatible with various types of images. Inspired by the medium transmission models [54]–[57] that characterize the light energy transmitted from the scene to the camera, we exploit a simple yet effective method to calculate pseudo depth from the distances between the surfaces of the scene points and the camera.

The *medium transmission* is the ratio of "the light energy that is not attenuated by the scene (including foreground and background) and reaches the observer" to "the real-world light energy reflected from the scene" [54]. According to the optics of the atmosphere, the attenuation is caused by the fact that particles along the transmission route always scatter the light, and "scattering" means that a particle absorbs a portion of the incident light and radiates the absorbed light as a light source [54]. Due to the layout of a scene, the beams of the

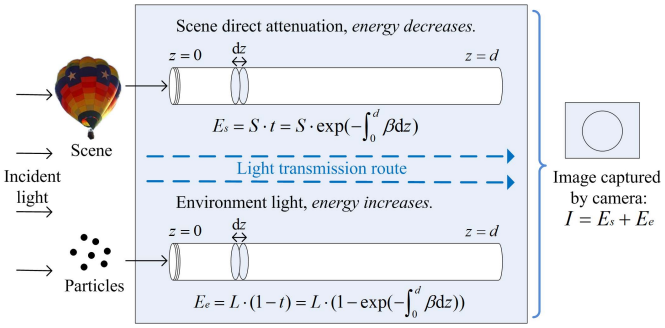

Fig. 1. Illustration of the medium transmission model. The total light energy captured by the camera is composed of the energies of the scene and the environment light. Please note that each beam denoted by a cylinder represents a unit area that corresponds to a pixel on an image, and the transmission map is spatially dependent.

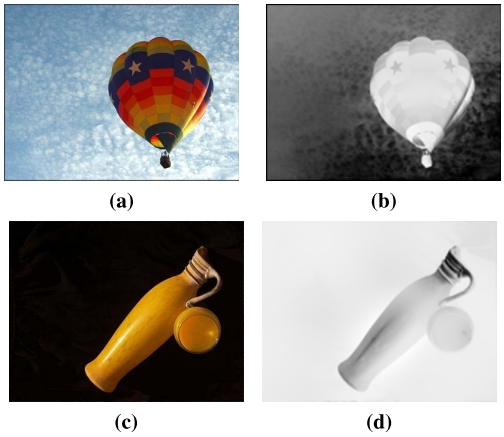

Fig. 2. Pseudo depth maps derived from [54]: (a) an outdoor image; (b) depth map of the outdoor image; (c) an indoor image; (d) depth map of the indoor image.

light from the surfaces of the scene points to the camera have different transmission routes, and the captured light energies on different positions of an image contain the distances of the transmission routes. That is, the medium transmission is spatially dependent.

Let (x, y) represent the spatial coordinates of a scene point, $t(x, y)$ measure the corresponding medium transmission, $S(x, y)$ be the real-world scene radiance. Then $S(x, y) \cdot t(x, y)$ measures "the light energy that is not attenuated by the scene and reaches the camera", which is also called *scene direct attenuation*. On the other hand, particles in the environment absorb and then radiate the light. This radiated light energy increases along the route of light transmission as the number of particles increases, and finally it also reaches the camera. Statistically, this *environment light* is spatially homogenous and can be regraded as a constant vector. Suppose the environment light is denoted by L, then $L \cdot (1 - t(x, y))$ measures the energy from environment light.

Given an RGB image I, the above medium transmission model is represented as

$$\begin{aligned} I(x, y) &= E_s(x, y) + E_e(x, y) \\ &= S(x, y) \cdot t(x, y) + L \cdot (1 - t(x, y)), \end{aligned} \tag{1}$$

where $I(x, y)$ represents an RGB vector at position (x, y), L is a constant vector, and E_s and E_e denote the energies from the scene and the environment respectively.

As shown in Fig. 1, the total light energy captured by the camera is composed of two parts, namely, the energy reflected by the scene and the environment light energy. These two parts cumulatively decrease and increase along the route of light transmission, respectively. That is to say, once the transmission map is obtained, we can calculate the depth value corresponding to this pixel. The estimated depth data is inverse proportional to the logarithm of the transmission as

$$d(x, y) = -\frac{\ln(t(x, y))}{\beta}, \tag{2}$$

where β is the scattering coefficient that is a spatial constant if the physical properties of the atmosphere are homogenous. Practically, we are more interested in the relative depth than the physical depth, and the real value of β is less important.

We apply the dark channel prior [54] to estimate the transmission map t according to Eq. (1). The dark channel prior is based on the statistics of normal images. It claims that, in most of the local regions of an image, there exist "dark pixels" that have very low intensities in at least one of the R, G, and B channels. According to [54], the transmission map at position (x, y) can be estimated by

$$t(x, y) = 1 - \phi \cdot \min_{\gamma} \left(\min_{\Omega} \frac{I(x, y)}{L} \right). \tag{3}$$

where $\gamma \in \{\mathrm{R,G,B}\}$ denotes the color channel; Ω represents pixels in a local region, e.g., a 15×15 patch around (x, y); ϕ is a decay parameter that allows slightly energy decrease on the captured image and it can be fixed at 0.95 as recommended in [54]. In practice, the constant vector L can be estimated from the pixel values corresponding to the dark pixels.

However, traditionally, the medium transmission models are always formulated for the application of outdoor scenes (such as the haze removal application [54], [55]), which are inapplicable to the images of indoor scenes. For example, as shown in Fig. 2, using the transmission model in [54], the pseudo depth map for the first image is sound, whilst the pseudo depth map of the second image is incorrect. This is because, for indoor scenes, the incident light may not come from the air but from the reflection of electric lights. To solve this problem, we consider the source of the incident light: if the incident light comes from the air light, we directly adopt the medium transmission model; when the incident light comes from the reflection of the electric lights, we reverse the strength of the captured images to simulate the real light energy. Based on this idea, we propose a robust pseudo depth measure termed *semi-inverse image depth*. Here, "semi-inverse image" means that we adaptively inverse the strength of the captured images according to the estimated light source to fit the real light transmission scenario of different types of images. Specifically, for an image I, we define its semi-inverse image as

$$\widehat{I}(x, y) = \begin{cases} 1 - I(x, y), & \text{if } \lambda \cdot LTN_c > LTN_s \\ I(x, y), & \text{otherwise,} \end{cases} \tag{4}$$

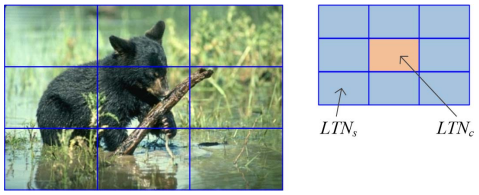

Fig. 3. Illustration of the lightness measure of the central and surrounding parts of an image.

where LTN_c and LTN_s represent the lightness of the central and surrounding parts of the image respectively and λ is an illumination coefficient. As illustrated in Fig. 3, the image is first divided into 3×3 blocks, and LTN_c and LTN_s are the mean lightness values of pixels in the central block and in the surrounding blocks, respectively. The coefficient λ is empirically set to 0.9.

Finally, the medium transmission model in Eq. (1) is adapted to

$$\widehat{I}(x, y) = S(x, y) \cdot t(x, y) + L \cdot (1 - t(x, y)), \tag{5}$$

and the pseudo depth can be estimated using Eqs. (4), (3) and (2) sequentially. In the following experiments, we normalize and invert the estimated depth to obtain the final pseudo depth for better visualization.

Provided with the consideration of the source of the incident light, our medium transmission model is suitable for different types of images. Besides, given an RGB image with $m * n$ pixels, the transmission map can be computed at the cost of the order $\mathcal{O}(mn)$, which is linear to the image size. Therefore, this model fulfills our requirement on efficiently extracting the pseudo depth of images. Fig. 4 shows a few examples of the semi-inverse image depth, from which the following observations can be made: (1) The pseudo depth fits both the outdoor and indoor scenes. It is more generic and robust than the depth map derived from the outdoor models, e.g., compare Figs. 4 (a) and (b) with Figs. 2 (b) and (d); (2) The pseudo depth provides object-level consistency of the salient objects, e.g., the humans in Figs. 4 (d) and (e). Note that, traditionally, it is hard to consistently extract the whole objects when parts of the objects are dissimilar in appearance; (3) It also deals well with the low brightness and the low contrast conditions, as shown in Figs. 4 (f) and (g); (4) When there exist cluttered environments (Figs. 4 (h) and (i)), the pseudo depth shows stably good performance. To summarize, the pseudo depth is very good at separating the foreground objects from the background and is hence useful for the task of salient object detection.

IV. RGB-'D' Saliency Detection

In this section, we propose a novel concept for salient object detection, named *RGB-'D' Saliency Detection*, by utilizing the pseudo depth derived on a single RGB image. We first discuss the conceptual definition and potential working directions in this new area, and then propose an algorithm for salient object detection following this concept.

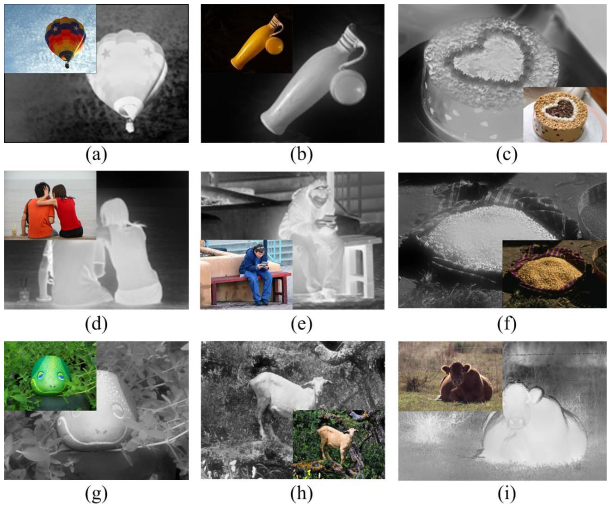

Fig. 4. Examples of the proposed semi-inverse image depth.

A. Conceptual Definition

The concept of RGB-'D' saliency detection is generally defined as: *given a 2D RGB image, deriving a pseudo depth cue on the image so as to make 2D to 3D conversion of the image, and then perform 3D saliency detection on the resulting pseudo 3D image.* There are four different perspectives to use the pseudo depth, which are summarized as follows and will be verified in the following studies of this paper.

1) Image Features: In salient object detection, an essential step is feature extraction on the input image. Traditional algorithms mainly adopt color and texture features. Now we can further use depth features, which naturally possess excellent discriminability for separating foreground objects and the background. We will illustrate this issue in Section IV-B-2).

2) Prior Knowledge: Most existing saliency detection algorithms use hand-crafted priors, such as the center prior [8], background prior [6], boundary prior [34], and focusness prior [11], which are shown to play crucial roles in enhancing the detection performance. Using pseudo depth, it is now possible to develop new and potentially better priors, such as the 'pseudo-depth-driven background prior' being described in Section IV-B-1).

3) An Additional Image Channel: Currently there exist a number of supervised salient object detection algorithms that directly learn saliency scores from raw features on RGB channels and exhibit good performance. The pseudo depth can be considered as an image channel in addition to the RGB channels. Then more discriminant features can be learned from the feature integration procedure. Based on this idea, an application will be given in Section VI.

4) Independent Depth-Induced Models: Deep features also show promising performance in saliency detection. We demonstrate that the pseudo depth can be used to generate independent depth-induced models and then to extract deep features for boosting the performance of saliency detection. An application will be provided in Section VI.

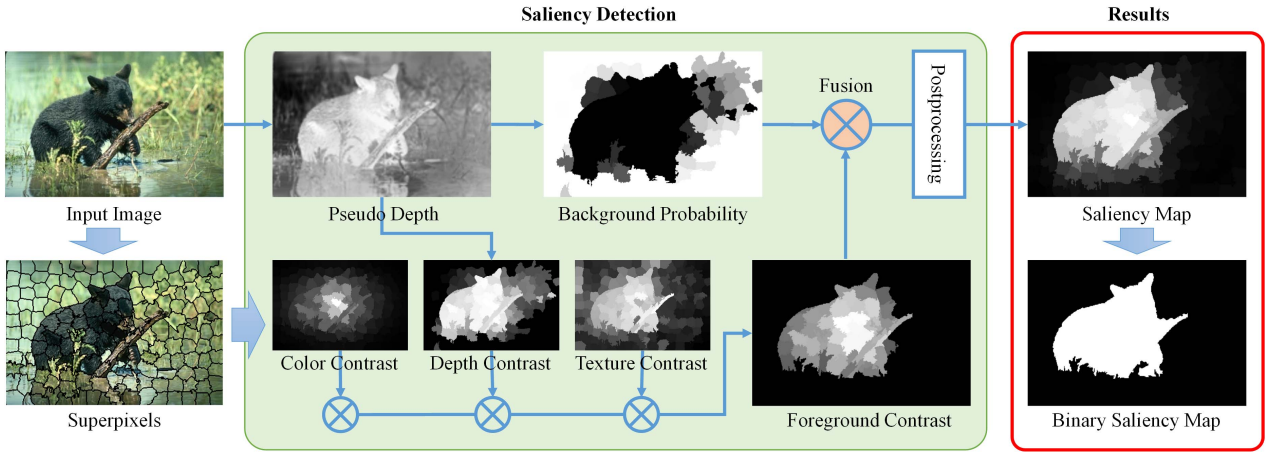

Fig. 5. The pipeline of the proposed PDP algorithm.

Note that RGB-'D' saliency detection is a general concept that it can be applied to develop different unsupervised and supervised algorithms. In this section, we first develop an unsupervised model with hand-crafted features and priors, so as to make it easier to observe and interpret the effectiveness of the RGB-'D' saliency concept. Further extensions with supervised methods are also feasible under this framework.

B. The Proposed Algorithm

As an application under the framework RGB-'D' saliency detection, we develop an unsupervised algorithm based on the Pseudo Depth Prior (PDP). The pipeline of PDP is shown in Fig. 5. The input image is segmented using superpixel segmentation algorithms [58]–[60] to generate the basic processing units. In our experiments, we adopt the simple linear iterative clustering (SLIC) algorithm [58] to segment the image into N superpixels due to its efficiency. For each superpixel, color, depth, and texture features are extracted so as to compute the respective contrast information. Aggregating the color, depth, and texture contrast measures, an initial foreground contrast map is obtained. On the other hand, we derive a background prior based on the proposed pseudo depth. The background prior and foreground contrast are then fused together via optimization. After a postprocessing step, the algorithm outputs the saliency map of the input image. We can also conduct a binarization step on the saliency map to obtain the binary version. The ingredients of PDP are detailed as follows.

1) Pseudo-Depth-Driven Background Prior: According to Borji *et al.*'s [1] comprehensive comparisons of 29 state-of-the-art salient object detection algorithms, all the top-performed algorithms explicitly utilize the background priors. The background priors possess good robustness in identifying the salient objects. In this work, we propose a novel background prior based on the pseudo depth.

As depicted in Fig. 4, the background regions of an image possess much lower depth values than the object regions. However, for saliency detection, some foreground objects with high depth values also do not grab human attention, which would better be classified into non-salient background regions. Figs. 4 (d) and (i) show some examples like the "stones" and the "grass". We can observe that this type of regions, such as stones and grass, always connect to the image borders. Zhu *et al.* [34] developed a boundary connectivity measure to separate the boundary regions. In this work, we adopt this boundary connectivity measure to refine our pseudo depth map. Let $\{d_i\}_{i=1}^N$ be the mean pseudo depth value of each superpixel and $\{bdCon_i\}_{i=1}^N$ be the boundary connectivity in [34], the refined depth $\{\hat{d}_i\}_{i=1}^N$ is calculated as

$$\hat{d}_i = \begin{cases} d_i, & \text{if } bdCon_i < \tau \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where the threshold τ is set to 2 as validated by the experiments in Section V-D1. The background strength of each superpixel is defined as the inverse of its refined depth value

$$bgStr_i = 1 - \hat{d}_i, \tag{7}$$

and the background probability of each superpixel is calculated as

$$P_i^{bg} = 1 - \exp\left(\frac{-bgStr_i^2}{2 \cdot \sigma_{bg}^2}\right), \tag{8}$$

where $i = 1, 2, \cdots, N$ and σ_{bg} is a bandwidth parameter that controls the variance of the generated probability values and is empirically set to 0.1. P_i^{bg} is close to 1 when the background strength is large and is close to 0 otherwise.

2) Foreground Contrast: Besides the background prior, we define a foreground saliency map by analyzing the contrast among superpixels, where three kinds of image features are adopted: color, depth, and texture.

a) Color contrast: We adopt the mean color value of each superpixel in the CIELAB color space since it is perceptually uniform with respect to human eyes. That is, the same amount of the numerical change in this space corresponds to (almost) the same amount of the visually perceived change [58]. The corresponding color distance between two superpixels is calculated as $dist^c(i, j) = ||c_i^{lab}, c_j^{lab}||_2$, where c_i^{lab} and c_j^{lab} are the color feature vectors of superpixels i and j, and

$||\cdot||_2$ represents the Euclidean distance. The color contrast map is calculated according to [7] and [12], by summarizing the spatially weighted color distances of one superpixel to all the other superpixels:

$$Ctr_i^c = \sum_{j=1}^{N} dist^c(i,j) \cdot \exp\left(\frac{-||p_i, p_j||_2^2}{2 \cdot \sigma_{spa}^2}\right), \tag{9}$$

where the weight is based on the distance between the center positions of superpixel i and j (i.e., $||p_i, p_j||_2$, and σ_{spa} is the bandwidth parameter of the spatial weighting scheme. The sensitivity of σ_{spa} is examined in Section V-D.1), and the recommended value for σ_{spa} is in $[0.2, 0.4]$.

b) Depth contrast: Using the pseudo depth of superpixels as features, the contrast is calculated based on the distance between the refined depth $\hat{d}_i$ and the minimum depth value of all superpixels ($\hat{d}_{\min} = \min\{\hat{d}_i\}_{i=1}^N$) as

$$Ctr_i^d = 1 - \exp\left(\frac{-||\hat{d}_i, \hat{d}_{\min}||_2^2}{2}\right), \tag{10}$$

where $\hat{d}_{\min}$ is assumed as the background depth. The larger distance from the superpixel depth to the background, the higher contrast value is assigned to the superpixel.

c) Texture contrast: For the texture feature, we use the differential excitation of the Weber Local Descriptor (WLD), which has shown to possess good discriminability and robustness [61]. Let h_i^{wld} and h_j^{wld} be the WLD histograms of superpixels i and j, the texture distance between the two superpixels is calculated using the chi-square distance as $dist^t(i,j) = \chi^2(h_i^{wld}, h_j^{wld}) = \sum_{b=1}^{bin} \frac{\left(h_i^{wld}(b) - h_j^{wld}(b)\right)^2}{h_i^{wld}(b) + h_j^{wld}(b)}$, where bin is the number of bins that represents the number of local texture patterns that are considered for contrast comparison. It is empirically set to 6. The texture contrast is then obtained by summarizing all texture distances of one superpixel to the others, weighted by their spatial distances:

$$Ctr_i^t = \sum_{j=1}^{N} dist^t(i,j) \cdot \exp\left(\frac{-||p_i, p_j||_2^2}{2 \cdot \sigma_{spa}^2}\right), \tag{11}$$

where the spatial weighting scheme is the same as that in Eq. (9).

The above color, depth, and texture contrast maps are visualized in Fig. 6, which capture different characteristic of the images. It can be observed that the depth contrast plays an important role in projecting the object level consistency of the salient objects. Then, the foreground contrast is calculated based on the Harmonic mean of the three contrast maps:

$$P_i^{fg} = \frac{3 \cdot (1 - P_i^{bg})}{(Ctr_i^c)^{-1} + (Ctr_i^d)^{-1} + (Ctr_i^t)^{-1}}, \tag{12}$$

where $(1 - P_i^{bg})$ is multiplied to enhance the accuracy of object identification (reduce the recall of background) and $i = 1, 2, \cdots, N$ is the index of superpixels.

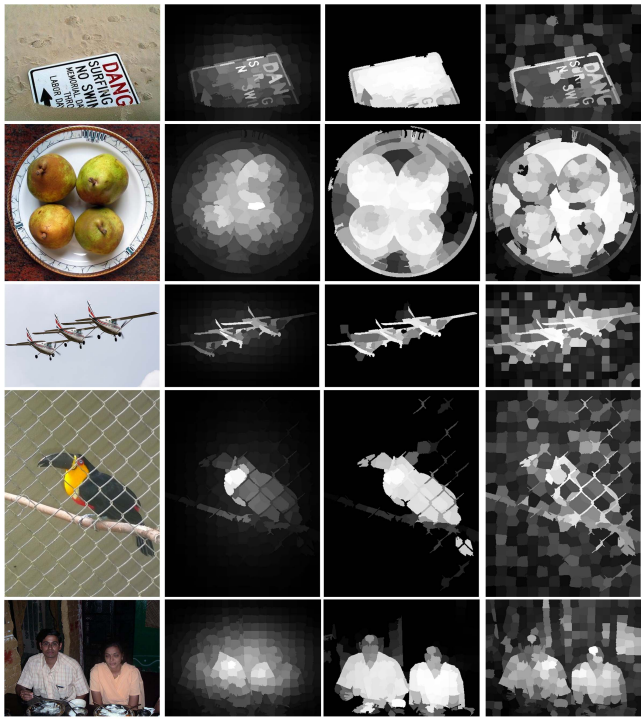

Fig. 6. Comparisons of different contrast maps: (a) input images; (b) color contrast; (c) depth contrast; (d) texture contrast.

3) Saliency Fusion: In the above procedures, two important cues for salient detection are derived, namely, the background probability map $\{P_i^{bg}\}_{i=1}^N$ and the foreground contrast map $\{P_i^{fg}\}_{i=1}^N$. Both maps are linearly normalized into the range of $[0, 1]$. When integrating these two maps, the final saliency map should be consistent with the data terms as well as being spatially smooth to enhance visual consistency. Suppose $\boldsymbol{f} = [f_1, f_2, \cdots, f_N]^T$ contains the optimal salient values of all superpixels in the image, it is obtained by minimizing the cost function

$$J = F_{data} + F_{smooth}, \tag{13}$$

where F_{data} and F_{smooth} represent the costs produced in the integration of the background and foreground cues and the enhancement of spatial smoothness, respectively. Specifically, they are defined as

$$F_{data} = \sum_{i=1}^{N} \left(\omega^{bg} \cdot P_i^{bg} \cdot f_i^2 + \omega^{fg} \cdot P_i^{fg} \cdot (1 - f_i)^2\right), \tag{14}$$

and

$$F_{smooth} = \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} \cdot (f_i - f_j)^2, \tag{15}$$

where ω^{bg}, ω^{bg}, and W are the fusion weights. To relieve the burden of manually balancing the data terms, we adaptively set the values of ω^{bg} and ω^{bg} according to their data uncertainty. In information theory, the larger the entropy is, the more

uncertainty the corresponding variable contains. To reduce the uncertainty of the data terms, we set their weights as

$$\omega^{bg} = \frac{1}{H_{bg}} \quad \text{and } \omega^{fg} = \frac{1}{H_{fg}}, \tag{16}$$

where $H_{bg} = -\sum_{i=1}^{N} P_i^{bg} \cdot \log_2 P_i^{bg}$ and $H_{fg} = -\sum_{i=1}^{N} P_i^{fg} \cdot \log_2 P_i^{fg}$. $\log_2 P_i^{fg}$ are the entropy of the background and foreground probabilities, respectively.

For the smoothness term, the weight matrix $\boldsymbol{W} = [W_{ij}]_{N\times N}$ is set according to the spatial distances among superpixels as

$$W_{ij} = \begin{cases} \exp\left(\dfrac{-||p_i, p_j||_2^2}{2 \cdot \sigma_{spa}^2}\right), & \text{if } i \text{ and } j \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

where σ_{spa} determines the strength of spatial smoothness, and this spatial weighting scheme is the same as those in Eqs. (9) and (11). Finally, the optimal superpixel-wise saliency values can be mathematically deduced as

$$\boldsymbol{f} = \arg\min_{\boldsymbol{f}} J = \frac{\omega^{fg} \cdot \boldsymbol{P}^{fg} \cdot \mathbb{1}^{\mathrm{T}}}{2 \cdot (\boldsymbol{D} - \boldsymbol{W}) + \omega^{bg} \cdot \boldsymbol{P}^{bg} + \omega^{fg} \cdot \boldsymbol{P}^{fg}}, \tag{18}$$

where

$$\begin{cases} \boldsymbol{P}^{bg} = \mathrm{diag}(P_1^{bg}, P_2^{bg} \cdots, P_N^{bg}) \\ \boldsymbol{P}^{fg} = \mathrm{diag}(P_1^{fg}, P_2^{fg} \cdots, P_N^{fg}) \\ \boldsymbol{D} = \mathrm{diag}(\sum_{j=1}^{N} W_{1j}, \cdots, \sum_{j=1}^{N} W_{Nj}.) \end{cases} \tag{19}$$

Afterwards, a fused saliency map S can be generated by filling the salient values of pixels in the ith superpixel with the value f_i.

4) Postprocessing: Based on the initial saliency map, PDP further performs a postprocessing step, mainly for the purpose of visualization enhancement. The procedure is described as follows.

Let τ be the Otsu's threshold [62] of the initial saliency map (S). We define a set of thresholds as $\{\lambda_i \cdot \tau\}_{i=1}^{K}$, where $\lambda_i = i/K$. For each threshold $\lambda_i \cdot \tau$, we set the saliency values that are smaller than this threshold to zero and keep the other saliency values as they are to generate a refined saliency map S_i, and thus K refined saliency maps $\{S_i\}_{i=1}^{K}$ are obtained in total. The optimized saliency map is then computed as

$$\widehat{S} = \frac{S + \sum_{i=1}^{K} \lambda_i \cdot S_i}{1 + \sum_{i=1}^{K} \lambda_i}. \tag{20}$$

Afterwards, $\widehat{S}$ is linearly normalized in the range of $[0, 1]$ as the final output. In our experiments, K is set to 5.

V. Experiments

To examine the performance of the proposed algorithm, we compare PDP with the state-of-the-art salient object detection methods.

A. Datasets and Competing Methods

Four standard datasets are used in the experiments: MSRA10K [19] includes 10,000 images, and many of them endure low contrast; ECSSD [37] contains 1000 natural images that are semantically meaningful but structurally complex; HKU-IS [63] is a newly opened dataset of 4447 natural images, while most of them have low contrast and contain multiple objects; and PASCAL-S [64] consists of 850 challenging images, whereas many images contain multiple objects in cluttered environment. We compare PDP with the state-of-the-arts, including 12 unsupervised algorithms (CA [15], DSG [18], DSR [13], GS [6], HS [37], MAP [65], MC [10], MR [9], MST [17], RBD [34], RC [12], SF [7]) and two supervised ones (BL [25] and MIL [27]).

B. Evaluation Metrics

As recommended in [1], we adopt the Precision-Recall (PR) curve, Area Under ROC Curve (AUC) score, Mean Absolute Error (MAE) score, F-measure, and weighted-F_β (w-F_β) to compare the performance of competing algorithms. To illustrate the evaluation metrics, in the following statement, S represents the detected saliency map and M is used to denote the binarized S using thresholds sliding from 0 to 1 with step 1/255; G represents the human-labeled binary ground truth; $|\cdot|$ stands for the number of pixels in current set, while $||\cdot||_1$ is the L_1 norm distance.

1) Precision-Recall: The precision and recall values are calculated as follows

$$Precision = \frac{|M \cap G|}{|M|}, \quad Recall = \frac{|M \cap G|}{|G|}. \tag{21}$$

2) F-Measure: To simultaneously consider the precision and recall values, the F-measure is calculated as

$$F_\beta = \frac{(1 + \beta_f^2) \cdot Precision \cdot Recall}{\beta_f^2 \cdot Precision + Recall}, \tag{22}$$

where $\beta_f^2 = 0.3$ is used to magnify the effect of the precision scores [1].

3) Area Under ROC Curve (AUC) Score: To evaluate the predicted saliency maps, the Receiver Operating Characteristics (ROC) curve explores the relationship between the true positive rate and false positive rate of these saliency maps. Then AUC calculates the area under the ROC curve and then concentrates this information into a single score. Higher AUC scores indicate better performance.

4) Mean Absolute Error (MAE) Score: The MAE score directly calculates the mean absolute distance between the predicted saliency map and the ground truth. It is calculated using $||S - G||_1$.

5) Weighted F-Measure: To explicitly rank different models, we also adopt the weighted-F_β (w-F_β) score, which is more reliable to evaluate the quality of saliency maps, eliminating the flaws (interpolation, dependency and equal-importance) of previous metrics like F_β and AUC. The detailed formulation of w-F_β is provided in [66].

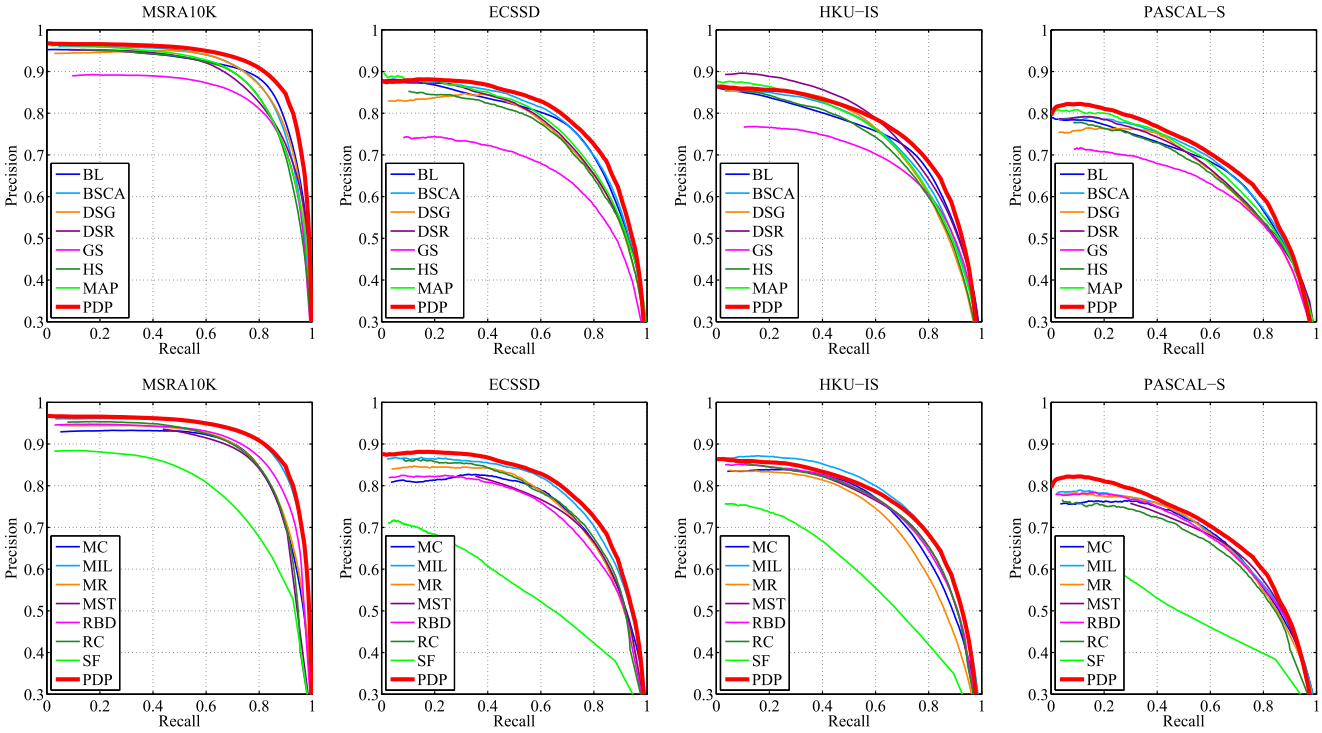

Fig. 7. PR curves of the competing algorithms.

TABLE I

F_β, w-F_β, AUC, AND MAE SCORES OF THE COMPETING ALGORITHMS

	MSRA10K				ECSSD				HKU-IS				PASCAL-S			
	F_β	w-F_β	AUC	MAE	F_β	w-F_β	AUC	MAE	F_β	w-F_β	AUC	MAE	F_β	w-F_β	AUC	MAE
BL	0.8632	0.7981	0.9636	0.1586	0.7544	0.6562	0.9157	0.2169	0.7205	0.6226	0.9158	0.2077	0.6628	0.5459	0.8694	0.2493
BSCA	0.8587	0.7738	0.9582	0.1252	0.7577	0.6421	0.922	0.1824	0.719	0.6025	0.9099	0.1748	0.6694	0.5264	0.8713	0.2238
DSG	0.8577	0.6766	0.9572	0.123	0.7312	0.511	0.8983	0.182	0.7194	0.4887	0.8884	0.1596	0.6505	0.4108	0.8384	0.2186
DSR	0.8345	0.7082	0.9589	0.1208	0.7346	0.5628	0.916	0.1782	0.7348	0.58	0.9232	0.1421	0.6517	0.4508	0.871	0.2081
GS	0.8141	0.7103	0.9476	0.1385	0.6607	0.5277	0.8839	0.2058	0.6771	0.5673	0.9104	0.1681	0.6239	0.479	0.859	0.2237
HS	0.8448	0.7511	0.9325	0.1486	0.7267	0.6045	0.8852	0.2274	0.704	0.581	0.8794	0.215	0.6451	0.5031	0.8378	0.2637
MAP	0.8449	0.7135	0.9517	0.1273	0.7401	0.5864	0.9055	0.1844	0.7164	0.5408	0.8917	0.1706	0.6613	0.475	0.8533	0.2242
MC	0.8475	0.7205	0.9507	0.1452	0.7394	0.575	0.9111	0.2023	0.7236	0.5571	0.9064	0.184	0.6675	0.4738	0.8714	0.2317
MIL	0.8805	0.7759	0.9707	0.1101	0.7602	0.6165	0.9186	0.177	0.7445	0.6031	0.9123	0.1576	0.6656	0.4984	0.8654	0.2199
MR	0.8463	0.7171	0.943	0.1265	0.7358	0.5726	0.8879	0.1892	0.7064	0.541	0.8663	0.1781	0.6647	0.4859	0.852	0.2231
MST	0.8413	0.7519	0.9209	0.0947	0.7244	0.6252	0.8939	0.1554	0.7216	0.6242	0.8942	0.1276	0.661	0.5462	0.8569	0.2035
RBD	0.8557	0.7561	0.9547	0.108	0.7164	0.5706	0.896	0.1752	0.723	0.5947	0.9081	0.1424	0.6589	0.5036	0.8655	0.2039
RC	0.8452	0.7455	0.9355	0.1373	0.7383	0.6071	0.8932	0.1859	0.7246	0.6046	0.901	0.1654	0.6473	0.5052	0.8403	0.2268
SF	0.7486	0.3997	0.9034	0.1708	0.5479	0.2253	0.7935	0.2187	0.5839	0.2528	0.8242	0.1744	0.4956	0.1582	0.7444	0.2412
PDP	0.8808	0.8241	0.9712	0.104	0.7677	0.6682	0.9238	0.1724	0.7368	0.6387	0.9164	0.1603	0.6765	0.5633	0.8803	0.2019

Red number indicates the best performance; blue number represents the second best performance.

C. Comparisons With Peer Models

1) Quantitative Evaluation: For the competing algorithms, we plot their PR curves in Fig. 7. The F_β, w-F_β, AUC, and MAE scores of different algorithms are reported in Table I. Generally speaking, PDP outperforms or is comparable to the state-of-the-arts.

More specifically, on the MSRA10K and ECSSD databases, PDP has the best performance considering the the PR curve, F_β, w-F_β and AUC scores; it also obtains the second lowest MAE scores. On the HKU-IS database, the PR curve of PDP is comparable to that of MIL. It ranks the second considering the F_β and AUC scores. Compared with the aforementioned evaluation metrics, the advantage of PDP on the MAE score is not so significant. On the HKU-IS database, PDP is outperformed by some algorithms (e.g., MST and DSR) considering the MAE score. This is because, the pixel-based methods (MST and DSR) have natural advantages over the superpixel-based methods (e.g., PDP) in measuring the pixel-wise absolute distance. The HKU-IS database has relatively complicated

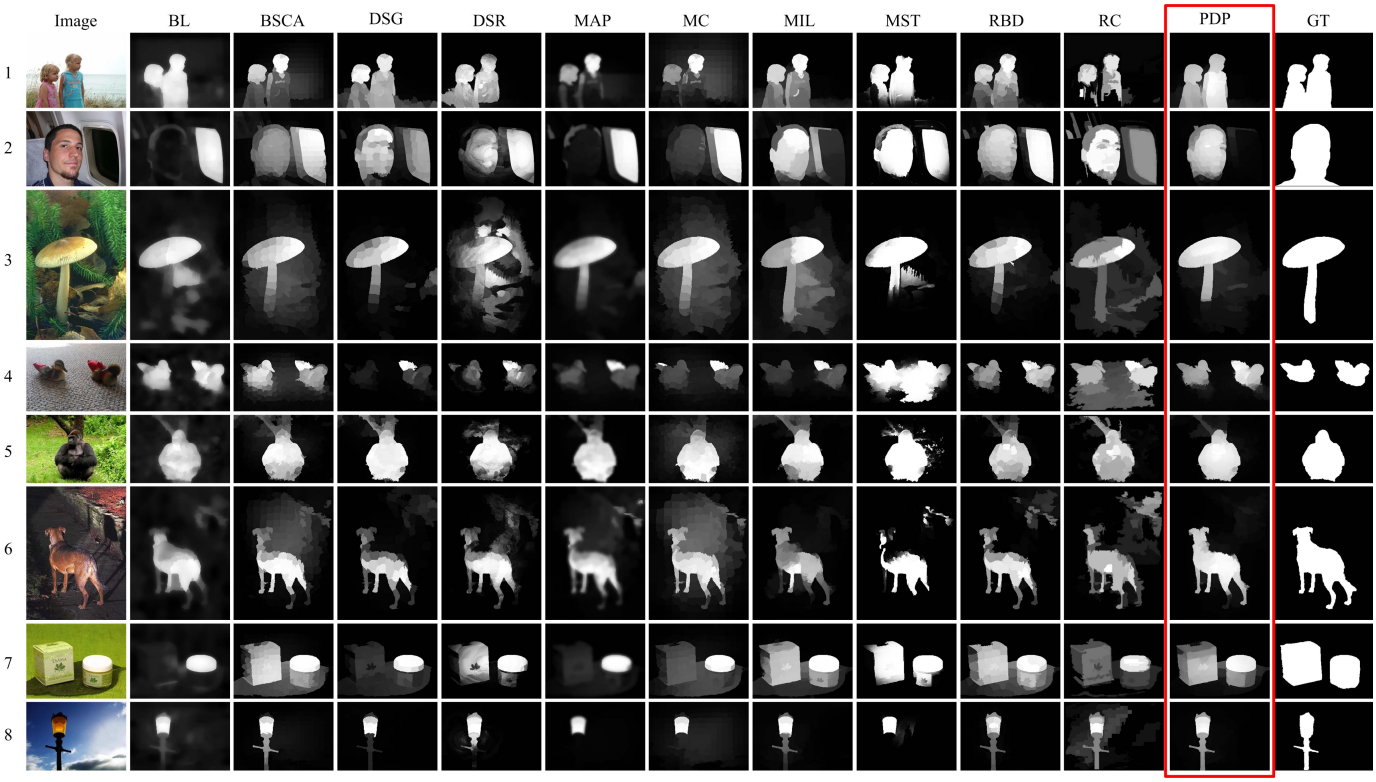

Fig. 8. Visual comparisons of the competing algorithms.

scene structures that degrade the MAE score of the saliency maps produced by the superpixel-based methods like PDP. Even though, PDP obtains the best w-F_β value on HKU-IS, showing that PDP is still competitive among the compared algorithms on this database. For the PASCAL-S database, PDP has the best performance according to all evaluation metrics.

Please note that the algorithm structure of PDP is similar to that of the RBD algorithm. Both of them adopt a background prior and the global contrast information. Compared with RBD, the superior performance of PDP validates the effectiveness of the RGB-'D' saliency concept.

2) Visual Comparisons: To present qualitative comparisons, several saliency maps of different algorithms are shown in Fig. 8. PDP obtains better visual results than the others, especially for the instances endure low color contrast (e.g., Image 5), cluttered or dark environment (e.g., Images 1, 3, and 6), object shadow (e.g., Image 4), distracting components (e.g., the window in Image 2), single object with different components (e.g., Image 8) and multiple objects (e.g., Images 1, 4, and 7). The good results owe much to the use of pseudo depth, which is resistant to the aforementioned situations.

D. Discussion

To provide a comprehensive understanding of the proposed algorithm, this section experimentally examines the properties of PDP considering (1) the sensitivity of model parameters; (2) the effectiveness of the pseudo depth cue; (3) the limitation and failure cases, and (4) the computation cost.

1) Sensitivity of Parameters: In this section, we examine the performance of PDP over the selection of model parameters. Experiments are conducted on the ECSSD database and comparisons are made among the corresponding PR curves. Specifically, we first investigate the performance of the background probability map over the threshold of the boundary connectivity (τ in Eq. (6)); then we examine the sensitivity of the bandwidth parameters in calculating the background probability map (σ_{bg} in Eq. (8)) and the color (σ_{spa} in Eq. (9)) contrast map. Please note that the spatial weighting schemes in the texture contrast (σ_{spa} in Eq. (11)) and the smoothness term (σ_{spa} (in Eq. (17))) are the same as that in calculating the color contrast, and they have similar performance over the selection of σ_{spa}. For simplicity, these three parameters are set to the same value 0.4 in all experiments. The PR curves of the feature maps under the corresponding parameter settings are shown in Fig. 9. As can been seen, (1) the background probability map is insensitive to the threshold of the boundary connectivity for $\tau = 2, 3, 4, 5$; (2) the background probability map has stable performance when $\sigma_{bg} \in [0.05, 0.8]$; (3) the color contrast obtains better performance with a moderate spatial weighting scheme, i.e., $\sigma_{spa} \in [0.2, 0.4]$. To summarize, the performance of PDP is influenced by the choices of the aforementioned model parameters, however, it is not highly sensitive to them.

2) Validation of the Effectiveness of the Pseudo Depth: The pseudo depth cue provides rich information on the 3D cues of the scene to assist saliency detection. As it is involved in calculating the background probability and the foreground contrast, we investigate the effectiveness of the pseudo depth cue in these two parts, respectively. For the background probability, we compare our pseudo-depth-driven background

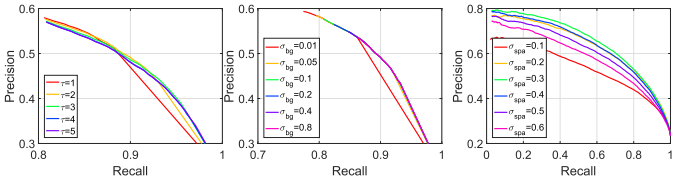

Fig. 9. Sensitivity analysis of model parameters on the ECSSD database: PR curves of the feature maps with different values of (a) the threshold τ of the boundary connectivity; (b) the bandwidth parameter σ_{bg} in the background probability map; (c) the bandwidth parameter σ_{spa} in the color contrast map.

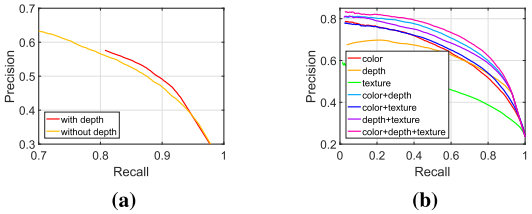

Fig. 10. Validation of the effectiveness of the pseudo depth on the ECSSD database: PR curves of (a) the background probability maps with and without pseudo depth; (b) different combinations of the contrast maps.

prior with the background connectivity prior without pseudo depth; in terms of the foreground contrast, we examine the performance of the color, depth, and texture contrasts and their different combinations. The results in Fig. 10 demonstrate the beneficial effects of the pseudo depth.

3) Failure Cases: As our algorithm adopts the medium transmission model to generate the pseudo depth, it may become less effective when the medium transmission model fails in some cases. Fig. 11 presents some failure examples. For Image 1, the fur of the cat has poor reflection ability compared with the blanket, and the inferiority is further aggravated by the large area of the blanket. For Image 2, the multiple light sources bring the challenge of estimating the light transmission route, and hence the depth map is disordered. Besides, it is also hard to generate a reliable pseudo depth from extremely cluttered scenes, as illustrated in Image 3. Please note that, under the RGB-'D' saliency framework, we can resort to an advanced depth estimation model to solve the problems in the first two images since the medium transmission model is easily affected by the refection of the surfaces of objects.

4) Time Analysis: The computation cost of PDP is linear to the size of the input image. We implemented it using MATLAB and run it on a platform with i7 3.4 GHz CPU and 16GB RAM. A single CPU is used in the execution. For images with 400×300 pixels, PDP requires about 900 milliseconds for processing, among which the calculation of the pseudo depth costs 85 milliseconds; superpixel segmentation costs 96 milliseconds; feature extraction needs 699 milliseconds; the calculation of background probability and foreground contrast information uses 15 and 1 milliseconds, respectively; the saliency fusion step requires 3 milliseconds; and post-processing requires 1 millisecond. It can be observed that most of the computational overhead of PDP comes from the feature extraction step, while all the other procedures are very efficient.

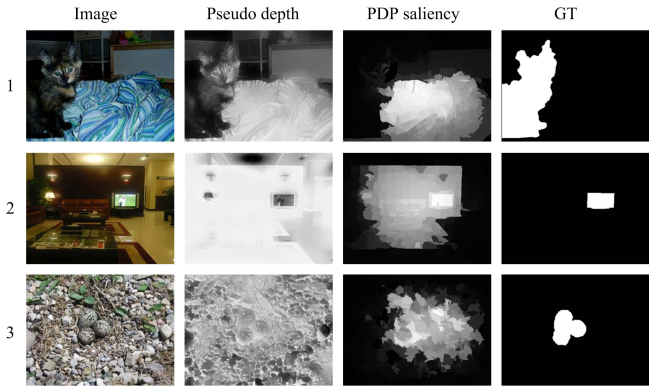

Fig. 11. Failure cases of PDP.

VI. Further Validation of the RGB-'D' Saliency Framework

As verified in Section V-C, PDP outperforms or is comparable to the state-of-the-arts. We use this simple model with hand-crafted features to demonstrate and interpret the effectiveness of the proposed RGB-'D' saliency concept. To further validate the generalization ability of this framework, in this section, we adapt two newly developed RGB saliency models to our RGB-'D' saliency framework for potentially performance enhancement. A supervised feature integration method (DRFI [20]) and a method based on DNN features (AMC [35]) are used as two applications.

A. Integration With Existing Models

DRFI provides a computational model for the supervised feature-integration-based saliency algorithms. It learns a random forest regressor to integrate features for saliency prediction in a discriminative way. DRFI first extracts hand-crafted features from RGB channels, and then uses these features to train the saliency predictor. To integrate DRFI with the RGB-'D' saliency framework, we calculate the pseudo depth and consider it as an additional image channel, and then extract features from this pseudo depth channel following the processing of the RGB channels. More specifically, we concatenate the raw depth values, textures of depth, texture histograms of depth, and the local binary patterns of depth into the original RGB features and then train a new saliency predictor based on these RGB-'D' features. The adaption of DRFI to its RGB-'D' counterpart validates that the pseudo depth can work as an additional image channel when integrates with existing RGB saliency models.

AMC is based on the absorbing Markov chain of the image graph. Taking the absorbing time from superpixels to the image boundary nodes as their saliency values, the performance of this method relies heavily on the transition probability matrix, which encodes the similarities and dissimilarities among image superpixels. To provide a precise similarity measure, AMC exploits the deep color features

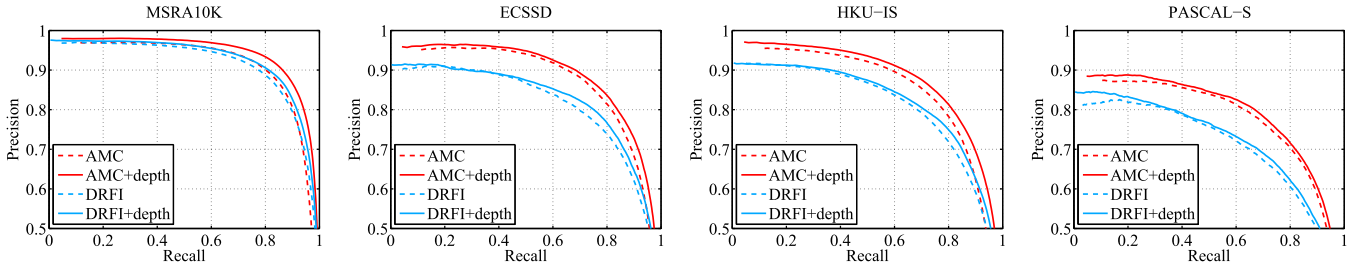

Fig. 12. PR curves of two algorithms and their adaptions to the RGB-'D' saliency framework.

TABLE II

F_β, w-F_β, AUC, AND MAE SCORES OF TWO ALGORITHMS AND THEIR ADAPTIONS TO THE RGB-'D' SALIENCY FRAMEWORK

	MSRA10K				**ECSSD**				**HKU-IS**				**PASCAL-S**			
	F_β	w-F_β	AUC	MAE	F_β	w-F_β	AUC	MAE	F_β	w-F_β	AUC	MAE	F_β	w-F_β	AUC	MAE
AMC	0.8778	0.7775	0.9519	0.1131	0.8321	0.7151	0.9247	0.152	0.8134	0.6916	0.9214	0.1337	0.7546	0.6276	0.8895	0.1961
AMC+depth	0.899	0.829	0.9758	0.1057	0.8393	0.7446	0.9491	0.1522	0.8278	0.738	0.9524	0.1383	0.7658	0.6488	0.9134	0.1975
DRFI	0.8682	0.7654	0.9714	0.131	0.7752	0.6237	0.9383	0.1719	0.7712	0.6419	0.9437	0.1457	0.6896	0.5047	0.8955	0.2111
DRFI+depth	0.8792	0.7856	0.9765	0.1164	0.7913	0.6422	0.9462	0.1598	0.7791	0.6533	0.9519	0.1343	0.6977	0.5193	0.903	0.2003

Red number indicates the better performance.

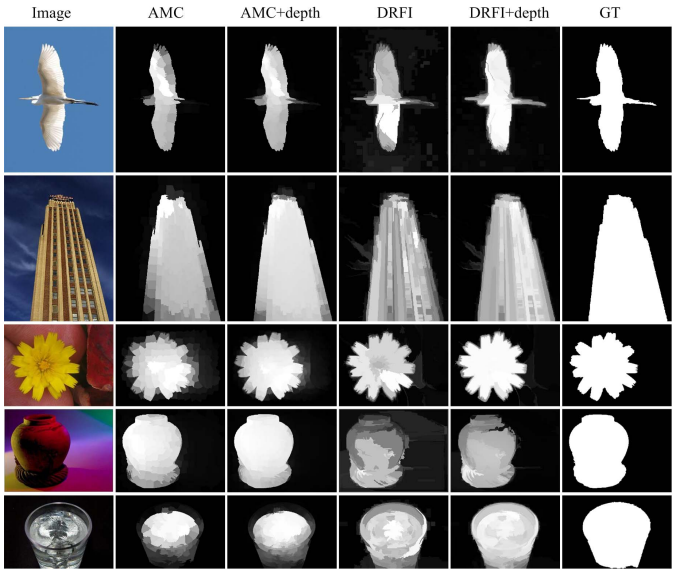

Fig. 13. Visual comparison of two algorithms and their adaptions to the RGB-'D' saliency framework.

extracted from the fully convolutional network, i.e., FCN-32s RGB [67], to learn the transition matrix. In this example, we adopt a pre-trained depth network, FCN-32s HHA [67], to obtain deep depth features. The HHA model is a three-dimensional geometric depth model generated from a single depth map. We use our pseudo depth map to generate a HHA model, and then input this model into the pre-trained FCN-32s HHA network to extract deep depth features from the images. Afterwards, the deep depth features can be plugged into the AMC framework to generate a new transition matrix and correspondingly a new saliency map. Finally, we fuse the depth-based saliency map with the original color-based saliency map by average to get the final saliency map. This example explores the performance of RGB-'D' saliency by exploiting the pseudo depth to generate independent depth-induced models.

B. Experimental Validation

To evaluate the performance of our RGB-'D' saliency framework, we conduct both quantitative and qualitative evaluations on the two models and their adaptions to the RGB-'D' saliency framework. As show in Fig. 12, both methods have significant improvements on their PR curves after adaption. They also have better F_β, w-F_β, AUC and MAE scores in most cases (Table II). The visual comparisons are provided in Fig. 13. As we can see, the usage of pseudo depth improves the object-level consistency of the salient objects and hence is beneficial for performance enhancement. To summarize, in addition to working as basic features or saliency priors, the pseudo depth can be considered as an additional image channel or independent depth-induced models to assist saliency detection, demonstrating the generalization ability of the proposed RGB-'D' saliency detection concept.

VII. CONCLUSION

This paper proposed a new concept named RGB-'D' saliency detection by exploiting the pseudo depth extracted from RGB-only images. Specifically, we developed a robust pseudo depth measure based on the semi-inverse of the RGB image, which has shown to be capable of perceiving the depth information of various types of images. Taking the pseudo depth as both a primary feature and the prior knowledge, we then developed an RGB-'D' saliency algorithm termed PDP. PDP is a straightforward and unsupervised approach, for the ease of observing the effectiveness of the RGB-'D' saliency concept. Experiments conducted on four large benchmark databases validated the promising performance of PDP.

Further, we adapted two existing RGB saliency models to the RGB-'D' saliency framework. The performance enhancement demonstrated the good generalization ability of the RGB-'D' saliency framework.

REFERENCES

[1] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[2] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. CVPR*, vol. 2, Jun. 2004, p. II–37.

[3] X. Liao *et al.*, "Automatic image segmentation using salient key point extraction and star shape prior," *Signal Process.*, vol. 105, pp. 122–136, Dec. 2014.

[4] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Proc. CVPR*, Jun. 2009, pp. 1007–1013.

[5] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[6] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, 2012, pp. 29–42.

[7] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. CVPR*, Jun. 2012, pp. 733–740.

[8] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. CVPR*, Jun. 2012, pp. 853–860.

[9] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, Jun. 2013, pp. 3166–3173.

[10] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. ICCV*, Dec. 2013, pp. 1665–1672.

[11] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by UFO: Uniqueness, focusness and objectness," in *Proc. ICCV*, Dec. 2013, pp. 1976–1983.

[12] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. 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.

[13] H. Lu, X. Li, L. Zhang, X. Ruan, and M. H. Yang, "Dense and sparse reconstruction error based saliency descriptor," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1592–1603, Apr. 2016.

[14] Q. Liu, X. Hong, B. Zou, J. Chen, Z. Chen, and G. Zhao, "Hierarchical contour closure-based holistic salient object detection," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4537–4552, Sep. 2017.

[15] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. CVPR*, Jun. 2015, pp. 110–119.

[16] Y. Qin, M. Feng, H. Lu, and G. W. Cottrell, "Hierarchical cellular automata for visual saliency," *Int. J. Comput. Vis.*, vol. 126, no. 7, pp. 751–770, 2018.

[17] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. CVPR*, Jun. 2016, pp. 2334–2342.

[18] L. Zhou, Z. Yang, Z. Zhou, and D. Hu, "Salient region detection using diffusion process on a two-layer sparse graph," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5882–5894, Dec. 2017.

[19] 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.

[20] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. CVPR*, Jun. 2013, pp. 2083–2090.

[21] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. CVPR*, Jun. 2016, pp. 478–487.

[22] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. CVPR*, Jun. 2016, pp. 678–686.

[23] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. ECCV*, 2016, pp. 825–841.

[24] X. Wang, H. Ma, X. Chen, and S. You, "Edge preserving and multi-scale contextual neural network for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 121–134, Jan. 2018.

[25] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. CVPR*, Jun. 2015, pp. 1884–1892.

[26] H. Lu, X. Zhang, J. Qi, N. Tong, X. Ruan, and M.-H. Yang, "Co-bootstrapping saliency," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 414–425, Jan. 2017.

[27] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1911–1922, Apr. 2017.

[28] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, 2014, pp. 92–109.

[29] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Y. Yang, "Exploiting global priors for RGB-D saliency detection," in *Proc. CVPR*, Jun. 2015, pp. 25–32.

[30] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. CVPR*, Jun. 2016, pp. 2343–2350.

[31] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 2017.

[32] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*. Springer, 2012, pp. 746–760.

[33] M. Stommel, M. Beetz, and W. Xu, "Inpainting of missing values in the Kinect sensor's depth maps based on background estimates," *IEEE Sensors J.*, vol. 14, no. 4, pp. 1107–1116, Apr. 2014.

[34] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. CVPR*, Jun. 2014, pp. 2814–2821.

[35] L. Zhang, J. Ai, B. Jiang, H. Lu, and X. Li, "Saliency detection via absorbing Markov chain with learnt transition probability," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 987–998, Feb. 2018.

[36] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[37] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. CVPR*, Jun. 2013, pp. 1155–1162.

[38] Q. Wang, W. Zheng, and R. Piramuthu, "GraB: Visual saliency via novel graph model and background priors," in *Proc. CVPR*, Jun. 2016, pp. 535–543.

[39] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. CVPR*, Jun. 2013, pp. 1131–1138.

[40] Y. Xu, X. Hong, F. Porikli, X. Liu, J. Chen, and G. Zhao, "Saliency integration: An arbitrator model," *IEEE Trans. Multimedia*, to be published.

[41] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6038–6051, Dec. 2018.

[42] L. Zhang, X. Fang, H. Bo, T. Wang, and H. Lu, "Deep multi-level networks with multi-task learning for saliency detection," *Neurocomputing*, vol. 312, pp. 229–238, Oct. 2018.

[43] 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.

[44] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1226–1238, Sep. 2002.

[45] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," in *Proc. CVPR*, Dec. 2001, p. 325.

[46] X. He and A. Yuille, "Occlusion boundary detection using pseudo-depth," in *Proc. ECCV*. Springer, 2010, pp. 539–552.

[47] [Online]. Available: https://github.com/Microsoft/AirSim

[48] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[49] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug. 1999.

[50] A. Saxena, S. H. Chung, and A. Y. Ng, "3-D depth reconstruction from a single still image," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 53–69, 2008.

[51] E. Delage, H. Lee, and A. Y. Ng, "A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2418–2428.

[52] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. NIPS*, 2006, pp. 1161–1168.

[53] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[54] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.
[55] F. Guo, J. Tang, and H. Peng, "Adaptive estimation of depth map for two-dimensional to three-dimensional stereoscopic conversion," *Opt. Rev.*, vol. 21, no. 1, pp. 60–73, 2014.
[56] J.-M. Guo, J.-Y. Syue, V. R. Radzicki, and H. Lee, "An efficient fusion-based defogging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4217–4228, Sep. 2017.
[57] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2856–2868, Jun. 2018.
[58] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[59] Y.-J. Gong and Y. Zhou, "Differential evolutionary superpixel segmentation," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1390–1404, Mar. 2018.
[60] X. Xiao, Y. Zhou, and Y.-J. Gong, "Content-adaptive superpixel segmentation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2883–2896, Jun. 2018.
[61] J. Chen *et al.*, "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, Sep. 2010.
[62] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[63] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. CVPR*, Jun. 2015, pp. 5455–5463.
[64] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. CVPR*, Jun. 2014, pp. 280–287.
[65] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on Markov absorption probabilities," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1639–1649, May 2015.
[66] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. CVPR*, Jun. 2014, pp. 248–255.
[67] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

Xiaolin Xiao received the B.E. degree in software engineering from Wuhan University, China, in 2013. She is currently pursuing the Ph.D. degree with the Department of Computer and Information Science, University of Macau, Macau, China. Her research interests include superpixel segmentation, saliency detection, and color image processing and understanding.

Yicong Zhou (M'07–SM'14) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA.

He is currently an Associate Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include chaotic systems, multimedia security, computer vision, and machine learning.

Dr. Zhou was a recipient of the Third Price of Macau Natural Science Award in 2014. He serves as an Associate Editor for *Neurocomputing*, *Journal of Visual Communication and Image Representation*, and *Signal Processing: Image Communication*. He is a Co-Chair of the Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He is a Senior Member of the International Society for Optical Engineering (SPIE).

Yue-Jiao Gong (M'15) received the B.S. and Ph.D. degrees in computer science from Sun Yat-sen University, China, in 2010 and 2014, respectively. During 2015–2016, she was a Post-Doctoral Research Fellow with the Department of Computer and Information Science, University of Macau, Macau. She is currently an Associate Professor with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include evolutionary computation and machine learning methods, and their applications to image processing. She has published over 50 papers in the above-mentioned fields. She currently serves as a Reviewer for IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, IEEE TRANSACTIONS ON NEURAL NETWORK AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON IMAGE PROCESSING.