Multimedia Tools and Applications (2021) 80:16329–16346
https://doi.org/10.1007/s11042-020-08890-x

Region-based depth feature descriptor for saliency detection on light field

Xue Wang[1] · Yingying Dong[1] · Qi Zhang[1] · Qing Wang[1]

Received: 4 June 2019 / Revised: 3 January 2020 / Accepted: 27 March 2020 /
Published online: 28 April 2020

Abstract

This paper addresses the light field saliency detection problem via a multiple cue integration framework. By reinterpreting the usage of dark channels in estimating the amount of defocus, a novel region-based depth feature descriptor (RDFD) defined over the focal stack is proposed. Compared to the methods which utilize the depth map as another image channel, the RDFD can produce more informative saliency cues and make less restrictive assumptions on accurate depth map or focused clear images containing dark pixels. The proposed RDFD facilitates saliency detection in the following two respects: (1) the region-based depth contrast map can be computed by measuring a pair-wise distance between superpixels with the proposed RDFD, (2) a spatial distribution prior in the 3D space (3D-SDP) can be obtained from such depth measurements to provide high-level semantic guidances, including the gradient-like distribution in depth and the object-biased prior in image plane. Both of them contribute to generating a contrast-based depth saliency map and refining a background-based color saliency map. Finally, these saliency maps are merged into a single map using a multi-layer cellular antomata (MCA) optimizer. Experimental results demonstrate that our method outperforms state-of-the-art techniques on the challenging light field saliency detection benchmark LFSD.

Keywords Light field · Saliency detection · Multiple cue integration · Dark channels · Depth feature descriptor

1 Introduction

Visualsaliency detection aims at locating the pixels or regions in a scene which most catch human visual attention. It plays an important preprocessing role in numerous computer vision applications, such as image retargeting [9], object recognition [32], image retrieval

✉ Xue Wang
xwang@nwpu.edu.cn

[1] School of Computer Science, Northwestern Polytechnical University,
Xi'an 710072, People's Republic of China

[22, 23, 40], image compression [13] and image segmentation [10]. Many saliency models have been developed for predicting eye fixation [6, 14, 36] or detecting salient objects [1, 2, 7, 17, 24]. The former focuses on estimating the points in 2D or 3D at which people are looking, while the latter tends to highlight the complete salient object as well as to eliminate the background.

Existing salient object detection solutions have focused on exploring various saliency cues or priors. Numerous 2D visual features has been proposed, including low-level features such as color, intensity, orientation, and high-level semantic descriptors such as objectness [17, 25, 38]. The color contrast prior has been widely accepted and used in almost all saliency models [7, 17, 20, 30, 41, 42]. However, 2D saliency models may fail when the foreground has a similar appearance with the background, or the texture in the foreground/background is cluttered. To address this, new methods combined depth contrast prior derived from RGB-D data [28, 31] or light fields [20, 41, 42] with 2D saliency features are proposed to improve the performance. Multiple cue integration proves to be a promising strategy for visual saliency detection tasks in complex scenarios.

Inspired by the mechanism of human visual attention, in this article, we propose a computational model for saliency detection in light fields by combing low-level features and high-level features. The basic idea is based on two observations. First, human eyes can conduct dynamic refocusing over different depth layers. But sometimes it is difficult to find out the foreground slice candidates or background slices due to the absence of high-level knowledge. Second, since human visual system can rapidly identify the salient object and ignore the background, it tends to capture objects within certain layers instead of all depth layers. Useful depth measurements for saliency detection tasks should serve to highlight complete salient objects and meanwhile eliminate the background, rather than determining accurate depth for each pixel. Therefore, we start with the dark channel to analyze its relationship with defocus blur. Then a novel RDFD extracted from the light field stack is proposed to facilitate both low-level and high-lever cues. Based on the RDFD, a contrast-based depth saliency map is generated by jointly taking the depth contrast map and the 3D-SDP into consideration. The proposed 3D-SDP aims to distinguish complete salient objects from distractors with distinctive distribution in depth and in image plane. In addition, a background-based color saliency map is also constructed by adopting the RDFD into background selection. Finally, the contrast-based depth saliency map and the background-based color saliency map are merged into a single map using a MCA optimizer. The pipeline of the proposed model is shown in Fig. 1.

Our major contributions can be summarized as follows:

(1) The usage of dark channels is reinterpreted to introduce a novel implicit depth descriptor for the light field by estimating the amount of defocus.
(2) A novel region-based depth feature descriptor (RDFD) extracted from the light field focal stack is introduced to provide more informative and robust depth cues for saliency detection. Unlike the methods using the depth map as another image channel, the extracted RDFDs provide enough discriminative information between the foreground and the background with even moderate depth estimation.
(3) A 3D spatial distribution prior derived from depth measurements is proposed to refine saliency estimations. The proposed 3D-SDP contributes to highlighting the complete salient object and simultaneously eliminating the background in both depth space and 2D image space.

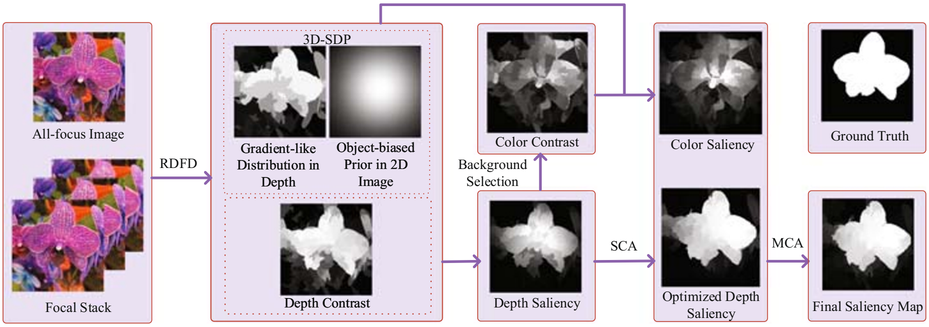


Fig. 1 The pipeline of the proposed light field saliency detection model

2 Related work

In past decades, extensive methods have been proposed for saliency detection. Readers can refer to [5] for comprehensive comparisons on state-of-the-art solutions. Here we only discuss most relevant methods from the following three aspects.

2.1 Spatial distribution prior

Since bottom-up approaches relying on low-level features tend to break a salient object into pieces [4], researchers propose to integrate high-level guidances into the saliency detection model, such as center-biased prior or object-biased prior. Center-biased prior assumes that salient objects are more likely to locate near the image center. Borji et al. [5] analyzed the influences of the center bias in model performances over challenging datasets. They found that a high precision can be easily achieved on images with large foreground objects by jointly considering the center bias property. Many existing solutions [11, 16, 33] have incorporated the center bias into their models, but the performance will be degraded when there is less center bias or small objects in images. Li et al. [21] used an object-biased Gaussian model as the 2D spatial distribution prior to refine saliency detection results. Our approach tries to combine the high-level guidance by proposing a 3D-SDP which includes both the gradient-like distribution in depth and the object-biased prior in image.

2.2 Background selection

Similar to the center-bias property of salient objects, background regions are prone to having the border-bias property. Background priors have been proved to be equally important, since one can eliminate the background to significantly improve foreground detection. Several approaches [15, 19, 21, 30, 39] considered that the image border is likely to be background. Li et al. [21] considered image boundaries as background templates and reconstructed the entire image by dense and sparse appearance models. Yang et al. [39] used the nodes on each side of an image as labelled background queries, then four side-specific maps can be generated. Jiang et al. [15] suggested that the random walk starting from background nodes could easily reach the absorbing nodes by using the boundary nodes as absorbing nodes. Qin et al. [30] applied the K-means algorithm to classify image borders into several clusters. Li et al. [19] constructed the non-saliency dictionary with super-pixels belonging to reference image boundaries and out-of-focus regions. However, Zhu et al. [43] found that it was fragile to treat all image boundaries as background when salient objects touched the boundary. Thus, they proposed another background measurement named boundary connectivity. They observed that an image patch belonged to the background only when the region was heavily connected to the image border.

Most of the above mentioned techniques rely on the color contrast. When the foreground and background have similar colors or textures, these approaches usually fail. Li et al. [20] calculated the focusness for each focal stack image and estimated depth layers with the assistance of underlying center prior. Background regions were selected with the focusness cue and the location cue. Zhang et al. [41] chose regions with higher background probability on the focusness map as the background. Our work exploits the depth cue embedded in the focal stack, which can significantly improve the saliency detection performance. Then regions with lower depth saliency values tend to be the background.

2.3 Depth cue

Depth cue has been proved to play an important role in saliency detection [18], which can be captured using a depth camera or estimated from stereo images. In the past decades, many studies have incorporated depth cue into their saliency detection models. Lang et al. [18] exploited the global-context depth prior extracted from the Kinect camera and estimated the joint density between saliency and depth using a mixture of Gaussians. Ciptadi et al. [8] explored the 3D layout and shape features from depth cues instead of simply treating depth as another channel of the input image.

Recently, several saliency detection models based on light fields have been proposed. Thanks to the efficient focusness and objectness cues embedding in a light field, these methods greatly improve saliency detection tasks in challenging scenarios such as similar foreground and background, cluttered background, and complex occlusions. Li et al. [20] pioneered a new saliency detection method on light field by utilizing the focusness map of each focal stack image to select the background and foreground candidates, which eliminated the limitation with depth maps. Zhang et al. [41] presented that saliency objects can be separated from the background by exploiting the depth-included contrast map. Sheng et al. [34] utilized the occlusion relationship to distinguish foreground and background regions. Besides, Li et al. [19] proposed a unified saliency detection framework for tailoring heterogenous types of input data, including 2D image data, 3D stereo data, and 4D light field data. To exploit the information embedded in 3D or 4D data, Zhang et al. [42] adopted the light-field flow for saliency detection. However, this method required discrete depth labels, which was very time-consuming. Sheng et al. [35] applied the inherent structure information in light field raw images for saliency detection.

Our approach resolves this issue by exploiting the depth cue embedded in a light field and defining a region-based depth feature descriptor RDFD. Based on the efficient and robust depth feature, we integrate depth contrast, objentness prior, color contrast and background prior together into a multiple cue integration framework.

3 Depth measurements

Before proceeding, we establish the following general notations. Let $\{I_m\}$, $m = 0, \ldots, M-1$ denote the focal stack synthesized from the light field and I^* denote the all-focus image by fusing the focused regions of $\{I_m\}$. We segment I^* into N small super-pixels using the simple linear iterative clustering (SLIC) algorithm [3]. Thanks to its ability of controlling the tradeoff between super-pixel compactness and boundary adherence, by adjusting the normalized color proximity parameter of SLIC, the resulting super-pixels would become smaller and have less regular size and shape. When the super-pixels are small enough, it is reasonable to assume that all the pixels belonging to one super-pixel would have similar depths. Such assumptions are used in many computer vision algorithms, such as window/segment-based techniques [26, 44], the edge consistency for stereo matching [37] and so on. More importantly, since we are more concerned with the depth-encoded feature distribution instead of accurate depth estimations, we adopt the depth hypothesis that all the pixels belonging to one super-pixel share the same depth as a simple means for more robustness. We use $\mathbf{x} = (x, y)$ and r to denote a pixel and a super-pixel respectively.

3.1 Defocus blur and dark channel

To motivate our work, we first describe the dark channel and its role in exploiting the depth cue embedded in the light field. For an image I, the dark channel [12] is defined by

$$D_I(\mathbf{x}) = \min_{\mathbf{x}' \in \Omega(\mathbf{x})} \left(\min_{c \in \{r,g,b\}} I^c(\mathbf{x}') \right), \tag{1}$$

where $\mathbf{x}$ and $\mathbf{x}'$ denote pixel locations, I^c is a color channel of I and $\Omega(\mathbf{x})$ is a local patch centered at $\mathbf{x}$. The defocus blur can be modeled as [29]

$$B = I \otimes k + n, \tag{2}$$

where B, I and n denote the blur image, latent image, and noise, respectively. We use the 2D Gaussian kernel $k(\mathbf{x}'|\mathbf{x}, \sigma^2) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$ as the point spread function (PSF). Note that we use $k(\sigma^2)$ to denote $k(\mathbf{x}'|\mathbf{x}, \sigma^2)$ in the following sections for simplicity. The standard deviation σ is proportional to the diameter of the circle of confusion (CoC), which can be used to measure the defocus blur.

The sparsity of dark channels has already been approved to be a natural metric to distinguish clear images from blurred images [27]. Furthermore, we find that the more blurred the image is, the fewer dark pixels it has, thus the dark channel can also be used to distinguish the degree of defocusing or blurriness. When the blurriness is uniform and spatial invariant, our observation can be formulated as

$$D_{B_1}(\mathbf{x}) \geq D_{B_2}(\mathbf{x}) \geq D_I(\mathbf{x}), \sigma_1 \geq \sigma_2, \tag{3}$$

where $B_1 = I \otimes k(\sigma_1^2) + n$, $B_2 = I \otimes k(\sigma_2^2) + n$. With the condition $\sigma_1 \geq \sigma_2$, the degree of blurriness of image B_1 is higher than that of B_2.

Proof Since $D_{B_1}(\mathbf{x}) \geq D_I(\mathbf{x})$ and $D_{B_2}(\mathbf{x}) \geq D_I(\mathbf{x})$ have already been approved in [27], we only need to determine the relationship between $D_{B_1}(\mathbf{x})$ and $D_{B_2}(\mathbf{x})$. Mathematically, based on the theory that the convolution of two Gaussian probability density functions is also a Gaussian, we have

$$\begin{aligned} B_1 &= I \otimes k\left(\sigma_1^2\right) + n \\ &= I \otimes k\left(\sigma_2^2\right) \otimes k\left(\sigma^2\right) + n \\ &= (I \otimes k(\sigma_2^2) + n - n) \otimes k(\sigma^2) + n \\ &= (B_2 - n) \otimes k(\sigma^2) + n \\ &= B_2 \otimes k(\sigma^2) - n \otimes k(\sigma^2) + n \\ &= B_2 \otimes k(\sigma^2) + n', \end{aligned} \tag{4}$$

where ${\sigma_1}^2 = {\sigma_2}^2 + \sigma^2$, $\sigma_1 > \sigma_2$. Thus we have $D_{B_1}(\mathbf{x}) \geq D_{B_2}(\mathbf{x})$ according to [27].

By taking the whole image into account, if there exists any pixel $\mathbf{x}$ satisfying $I(\mathbf{x}) = 0$, we have

$$\|D_{B_1}\|_0 \geq \|D_{B_2}\|_0 \geq \|D_I\|_0, \sigma_1 \geq \sigma_2, \tag{5}$$

where the L_0 norm $\|\cdot\|_0$ counts the number of non-zero entries in a vector or a signal.

3.2 Region-based depth feature descriptor

In order to extract the depth cue embedded in the focal stack at the super-pixel level, we propose a region-based depth feature descriptor by integrating the degree of defocusing over M focal stack images. The ideas behind the proposed RDFD originates from these two observations: (1) A set of focal stack images can be generated from a light filed by focusing at different depth levels. On each focal stack image, small regions/patches located at the same depth tend to have the same degree of defocusing. (2) The dark channel prior can be used to estimate the degree of defocusing or blurriness.

The main steps of computing the RDFDs of the super-pixels of I^* are shown in Fig. 2. We first calculate dark channels of the all-focus image I^* and its corresponding focal stack images $\{I_m\}$ using (1), which can be written as D_{I^*} and D_{I_m} respectively. Since the dark channel prior would has no effect for image deblurring if the clear image contains no dark (zero-intensity) pixels [27], we adopt the differential operation in this work to remove this limitation. Thus the difference image $\Delta_m(\mathbf{x})$ can be computed using

$$\Delta_m(\mathbf{x}) = D_{I_m}(\mathbf{x}) - D_{I^*}(\mathbf{x}), 1 \leq m \leq M. \tag{6}$$

Based on the assumption that a small region, such as a super-pixel, has the same depth and thus has the same degree of blurriness at a focal stack image, it is straightforward to define the M-dimensional RDFDs for each super-pixel as follows,

$$\mathbf{U}(r) = [U(r)_m]_{m=1}^{M}, U(r)_m = 1 - \frac{\|\Delta_m(r)\|_0}{T_r}, \tag{7}$$

where T_r is the total number of pixels belonging to the region/super-pixel r. The effectiveness of the differential operation for extracting the depth cue is shown in Fig. 3. The horizontal axis and the vertical axis represent focal stack image indexes and RDFD values respectively. Without the differential operation, for the labelled region which does not contain any dark pixels, its RDFDs over the focal stack display zero values and lose the discriminability. While with the differential operation, the distribution of RDFDs over the focal stack becomes distinguishable enough to extract the depth cue.

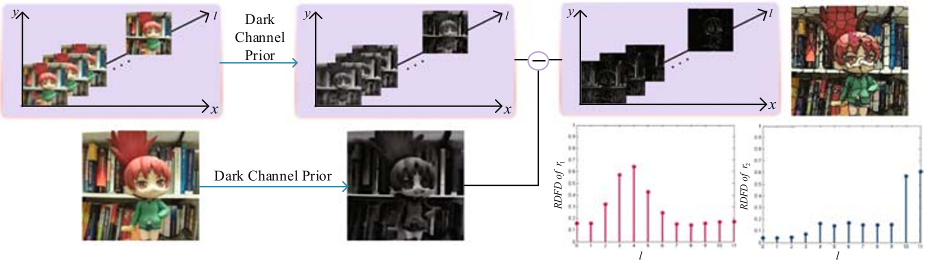
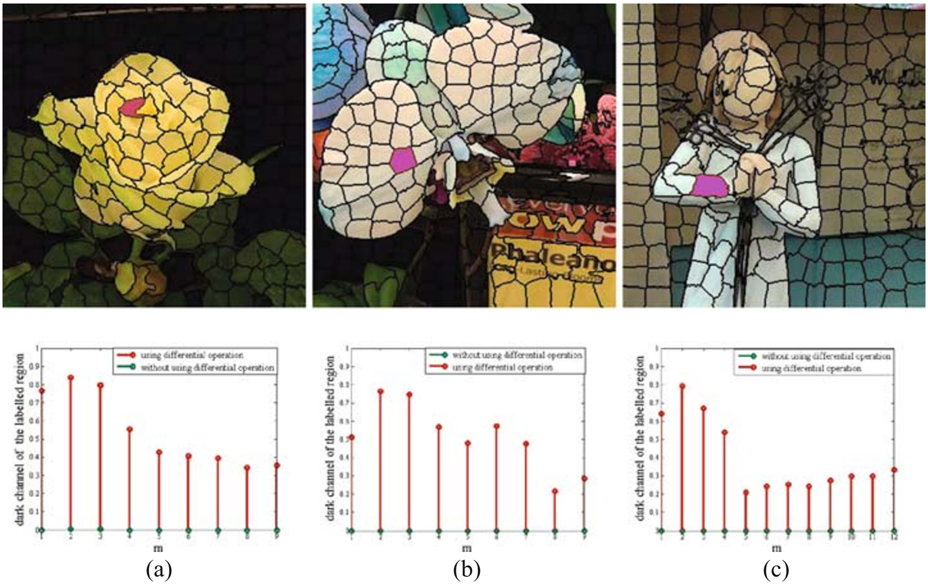


Fig. 2 The procedure of depth cue extraction. The foreground region r_1 focuses at I_4, and the background region r_2 focuses at I_{11}



(a) (b) (c)

Fig. 3 Effectiveness of the differential operation for extracting the depth cue, including the all-focus images on which the extracted boundaries from the over-segmentation and the interested super-pixels are superimposed (top) and corresponding RDFDs of the labelled super-pixels computed over the focal stack (bottom). The focused depth layers of the labelled super-pixels can be found by adopting the differential operation

Once we obtain the RDFDs of all the super-pixels, we are able to find the depth layer in a focal stack at which a super-pixel is maximally sharped by

$$\begin{aligned} l_r &= \arg\max \mathbf{U}(r) \\ &= \arg\min_{m} \frac{\|\Delta_m(r)\|_0}{T_r}, \end{aligned} \tag{8}$$

where l_r is the layer m in terms of depth. We call l_r the focused depth layer of super-pixel r.

4 Saliency detection model

In this section we estimate the saliency map for the all-focus image. Our saliency estimation is based on two sources of information, namely contrast-based depth saliency and background-based color saliency. Both benefit from the RDFD.

4.1 Contrast-based depth saliency

The proposed RDFDs generate more informative saliency cues in the following two respects: (1) the regional depth contrast map can be computed by measuring the pair-wise distances between super-pixels with the proposed RDFDs. (2) the 3D-SDP can be obtained from moderate depth measurements, including the gradient-like distribution in depth and the object-biased prior in the 2D image plane. Then, a contrast-based depth saliency map can be constructed by combining the depth contrast map and the 3D-SDP.

4.1.1 Spatial distribution prior in 3D

The 3D-SDP can be obtained from the proposed RDFD, including object-biased prior and a gradient-like distribution in depth. We use the proposed object-biased prior to refine our saliency performance instead of simply applying the image center assumption in 2D. In addition to the spatial information in 2D, experiments have demonstrated that human could rapidly direct attention to the attended depth plane. It is also proved that a gradient-like distribution in depth can be obtained with maximal processing at the attended depth plane, declining efficiency at more peripheral depths.

To extract the attended depth plane and potential salient object center, we apply K-means algorithm to divide N super-pixels into K sets, $\mathbf{S} = \left\{S^1, S^2, \ldots, S^K\right\}$ based on their focused depth layer and location information in the 2D image plane. Since a salient object should be complete and have a well-defined closed boundary [4], we set $K = 2$, referring to the foreground and background respectively. Given a set of observations $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N\}$, where each observation is defined by a three-dimensional real vector,

$$\mathbf{d}_r = [w_1 l_r, w_2 x_r, w_3 y_r], \tag{9}$$

where l_r is the focused depth layer of super-pixel r, (x_r, y_r) are the coordinates of the centroid of super-pixel r, w_1, w_2 and w_3 are the weights assigned to depth and position respectively. In the experiment we set $w_1 : w_2 : w_3 = 6 : 1 : 1$ to give a higher weight for the depth cue. After clustering, the k new centroids, one for each cluster, can be obtained, represented by $\mathbf{d}^k = \left[\omega_1 l^k, \omega_2 x^k, \omega_3 y^k\right]$, $1 \leq k \leq K$, where l^k and (x^k, y^k) correspond to the depth layer and the location of the centroid of the k-th cluster respectively. We compute the foreground cluster label f based on the observation that regions locating at the closer depth range tend to be the foreground

$$l^f = \arg\min_k \left\{l^k\right\}, \tag{10}$$

where l^f denotes the attended depth plane for the foreground. For each super-pixel, the gradient-like distribution in depth can be modeled as

$$\mathcal{L}_r = \exp(-\frac{(l_r - \mu)^2}{2{\sigma_3}^2}), \tag{11}$$

where $\mu = l^f$ and $\sigma_3 = 1/M$. We map the focused depth layer l_r of super-pixel r to $\mathcal{L}_r$ by adopting a Gaussian kernel, which achieves a maximum value at the attended depth plane. As a result, regions focused at the attended depth plane are assigned the maximum depth saliency value, while the regions focused at the farthest layer from the attended depth plane are assigned the minimum depth saliency value. The gradient-like distribution in depth makes it feasible to separate foreground from background once there exists differences between the foreground depths and the background depths. The mapping from l_r to $\mathcal{L}_r$ using a Gaussian kernel guarantees that foreground regions can still be assigned a higher value even when the estimation of the attended depth plane μ is inaccurate.

To render the salient object center instead of simply using the image center, we propose our object-biased model by

$$G(x, y) = \exp\left(-\left(\frac{(x - \mu_x)^2}{2{\sigma_x}^2} + \frac{\left(y - \mu_y\right)^2}{2{\sigma_y}^2}\right)\right), \tag{12}$$

where x and y are the coordinates of the pixel in I^*, μ_x and μ_y denote the coordinates of the foreground centroid. We set $\sigma_x = 0.25 \times W$ and $\sigma_y = 0.25 \times H$ as in [21], where W and H respectively denote the width and height of the all-focus image I^*.

Furthermore, we define the object-biased prior over the super-pixel using

$$G(r) = \frac{1}{T_r} \sum_{(x,y)\in r} G(x, y). \tag{13}$$

When the object does not locate at the image center, the proposed object-biased prior renders a more accurate object center (see Fig. 4), and therefore better refines the saliency detection results.

4.1.2 Depth saliency map

The similarity of any pair of super-pixels can be measured by the cosine distance with the proposed RDFD. We construct the depth contrast map by

$$DC(r_i) = \sum_{r_i \neq r_j} \exp\left(\frac{\left\|r_i, r_j\right\|}{-\sigma_4{}^2}\right) \cdot \left(1 - d\left(\mathbf{U}(r_i), \mathbf{U}\left(r_j\right)\right)\right), \tag{14}$$

where $\left\|r_i, r_j\right\|$ is the 2D Euclidean distance between the centroids of r_i and r_j, and $d\left(\mathbf{U}(r_i), \mathbf{U}\left(r_j\right)\right)$ is the cosine distance between the RDFDs of r_i and r_j. We set $\sigma_4{}^2 = 0.4$ with pixel coordinates normalized to [0, 1] as in [7].

We further construct the contrast-based depth saliency map by incorporating the proposed 3D-SDP to separate the foreground from the background,

$$DS(r) = (1 - \mathcal{L}_r) \cdot G(r) \cdot DC(r). \tag{15}$$

Inspired by [30], we use the Singly-layer Cellular Automata (SCA) to optimize the depth saliency map computed by (15). By considering a single super-pixel as a cell, SCA is capable of exploiting the internal relationship within the neighborhood of the cell. Different from [30], we make a modification to the original SCA method to improve the depth saliency

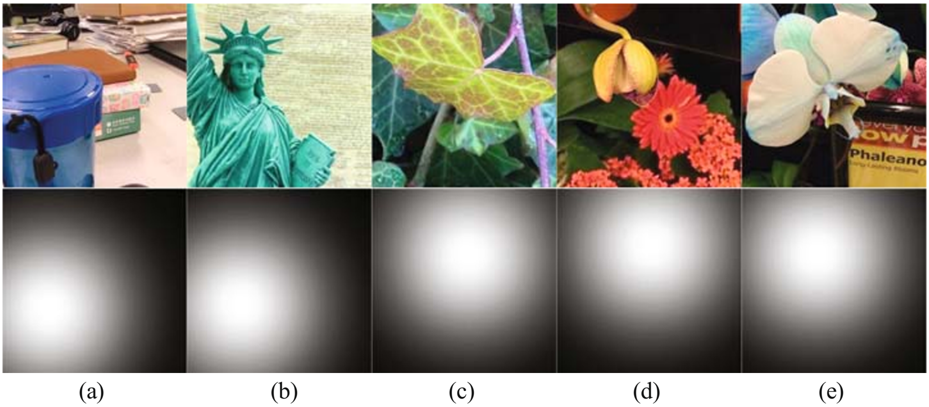


Fig. 4 Examples of SDPs in the 2D image space computed by our object-biased prior when the object does not locate at the image center

analysis. We construct the impact factor matrix $\mathbf{F} = \left[f\left(r_i, r_j\right)\right]_{N\times N}$ by defining the impact factor $f\left(r_i, r_j\right)$ for super-pixel r_i as

$$f\left(r_i, r_j\right) = \begin{cases} \exp\left(-\frac{1-d(\mathbf{U}(r_i),\mathbf{U}(r_j))}{\sigma_5{}^2}\right), & r_j \in NB(r_i) \\ 0, & \text{otherwise or} r_i = r_j \end{cases}, \tag{16}$$

where $NB\left(r_i\right)$ is the set of neighbors of r_i. We set $\sigma_5{}^2 = 0.1$ as in [30]. It makes sense that super-pixels within the same object tend to be homogeneous in perspective of depth. Thus, the neighbors with more similar depth features have greater influences on the next state of the cell. We denote the optimized depth saliency map as DS_{op}. Figure 5 shows the improvement of the optimized depth saliency map.

4.2 Background-based color saliency

The contrast-based depth saliency model can efficiently detect the salient foreground regions when the foreground and the background have different depth ranges. However, for the scenes of which the foreground and the background located at close depth layers or at the same depth range, the contrast-based depth saliency model would fail to separate the foreground from the background. Hence, we consider integrating the background-based color saliency model.

4.2.1 Background selection

Robust background prior can greatly improve the performance of saliency detection. Based on the observation that regions with lower depth saliency values are more likely to be backgrounds, we compute the background likelihood score $B(r)$ for each super-pixel by

$$B(r) = DS_{op}(r). \tag{17}$$

We further threshold the background likelihood score of all super-pixels for determining the background regions $\{B_r\}$, $1 \leq r \leq h$ in the all-focus image I^*, where h denotes the total number of super-pixels belonging to the background.

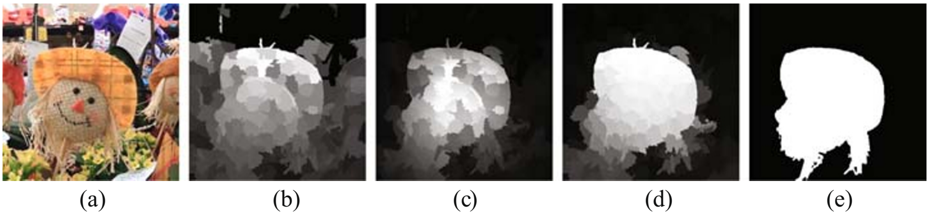


Fig. 5 The process of our depth saliency map construction model. (a) All-focus image. (b) Depth contrast map. (c) Depth saliency map by incorporating the 3D-SDP into the depth contrast map. Salient regions are highlighted and non-salient regions have been better suppressed. (d) Depth saliency map after optimization. Our modified SCA method generates a more smooth depth salient map. (e) Ground truth

4.2.2 Color saliency map

Once the super-pixels belonging to the background have been selected, the background-based color saliency map can be modeled as

$$CS(r_i) = G(r_i) \cdot \left[\frac{1}{h} \sum_{r_j \in \{B_r\}} \frac{1}{\max\limits_{c \in \{r,g,b\}} \left\| {r_i}^c, {r_j}^c \right\|} \right]^{-1}, \tag{18}$$

where $\left\| {r_i}^c, {r_j}^c \right\|$ is the Euclidean distance between the c-th color channels of r_i and r_j. The introduced object-biased term $G(r_i)$ contributes to the foreground enhancement as well as the background suppression, thus contributes to a better color saliency detection performance.

In order to combine the advantages of these two saliency maps DS_{op} and CS, we construct the final saliency map S by adopting the MCA optimization as in [30].

5 Experiments

We compare our approach with state-of-the-art methods including LFS [20], WSC [19], DILF [41] and MA [42]. The results of aforementioned methods are provided by their authors or found at corresponding project sites. The experiments are conducted on the Light Field Saliency Dataset (LFSD) [20]. Another dataset proposed for light field saliency detection, HFUT-Lytro dataset [42], is not used in our experiment, since the all-focus image is not spatially aligned with its focal stack images in terms of the image content, thus the differential operation in Eq.6 would cause errors. Theoretically, since the all-focus image and its corresponding focal stack images are extracted from the same raw image, the content of the all-focus image should be the same as that of its corresponding focal stack images. The focal stack images can be used to generate saliency cues or maps only when they have the same Field-of-View (FOV) with the corresponding all-focus image.

5.1 Experiment setup

Implementation details We set the number of super-pixels $N = 240$ in all the experiments, and the patch size Ω in Eq.1 as 7 × 7.

Evaluation metrics To conduct a quantitative performance evaluation, we use the precision-recall curve (PR curve), F-measure and the mean absolute error (MAE) to evaluate all the comparing saliency detection methods. Similar to previous works, we use a fixed threshold from 0 to 255 to binarize saliency maps. For each threshold, a pair of precision score and recall score is generated over the whole dataset.

The F-measure is a weighted harmonic mean of precision and recall defined by

$$F_\beta = \frac{\left(1 + \beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{19}$$

where β^2 is set to 0.3 to emphasize the precision term [2]. Here, we perform the evaluation using an adaptive threshold for each saliency map, which is defined the twice as much as the mean saliency of the saliency map S:

$$T = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y). \tag{20}$$

For a more comprehensive comparison, we compute the MAE between the continuous saliency map S and the binary ground truth annotation GT by

$$\mathrm{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \| S(x, y) - GT(x, y) \|. \tag{21}$$

5.2 Comparison with state-of-the-art methods

We first evaluate the performance of the proposed method against four state-of-arts methods, including LFS, WSC , DILF and MA. The quantitative evaluation results are shown in Fig. 6. Our approach significantly outperforms most comparing methods and achieves similar performance as DILF in terms of the precision-recall curve. In addition, the precision score of our algorithm is higher than that of DILF, only slightly below MA. One explanation for this is that the proposed depth measurements can significantly suppress the background, while the halo effect brought from the patch-based dark channel operation leads to a lower precision score. Moreover, our method achieves the best F-measure and the best MAE over state-of-the-art methods.

For visual comparisons, some examples are shown in Fig. 7, from which we can see that our method achieves a better background suppression performance, and generates saliency maps closer to the ground truth annotation.

In terms of computation complexity, we compare the average runtime for each sample among different light field saliency detection methods. We run the implementation by Matlab and C++ on an Intel i7 3.6GHz CPU PC with 16GB RAM. Table 1 shows the time cost of our approach compared with other state-of-the-art methods [19, 20, 41, 42]. It can be seen that our approach consumes a smaller amount of computing time than LFS, WSC and MA and a bit more than DILF.

5.3 Analysis of the proposed method

We further evaluate our method in details. In this paper, the most significant work is that we present a simple but effective method to extract depth cue embedded in the light field focal stack by using the dark channel prior. Hence we propose to use RDFD to construct

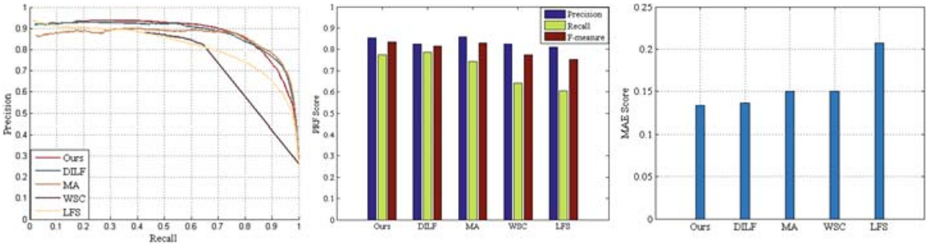


Fig. 6 Quantitative comparisons of different methods in LFSD. **Left**: precision-recall curves. **Middle**: precision, recall, and F-measure for adaptive thresholds. **Right**: MAE scores

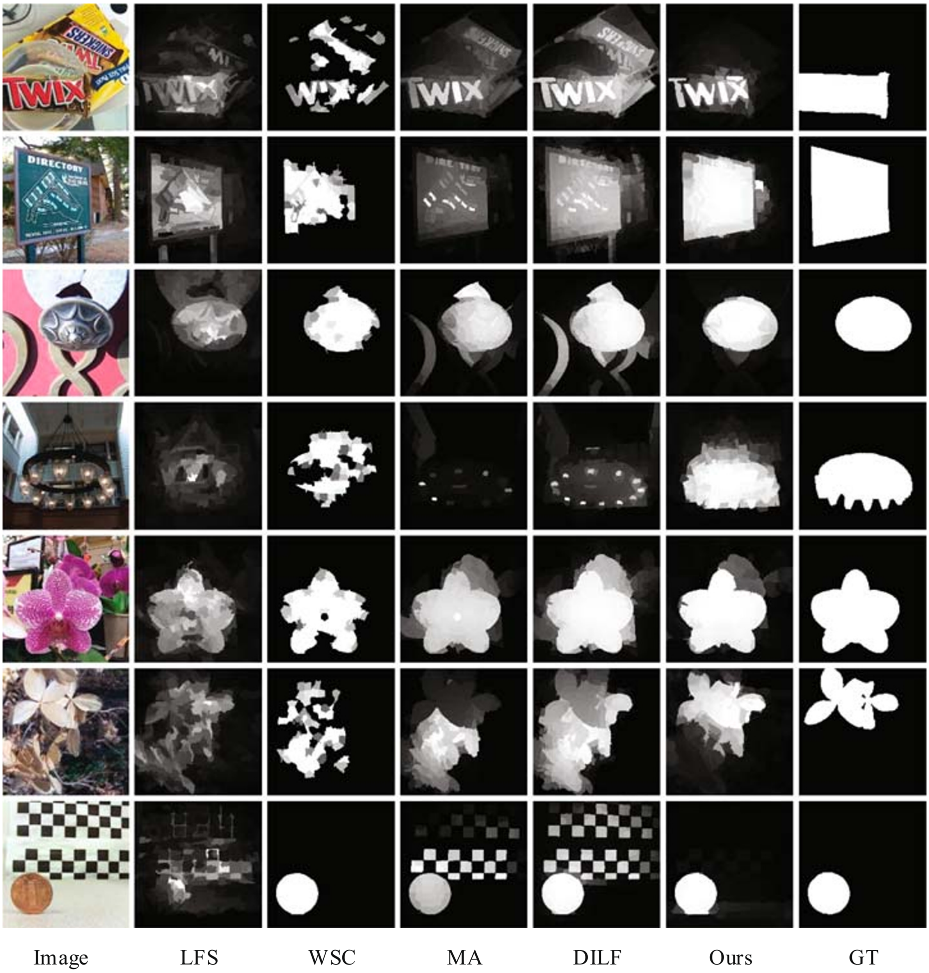


Fig. 7 Qualitative comparisons of different methods in LFSD. Our method generates saliency maps closer to the ground truth due to a better performance of foreground enhancement as well as background suppression

the contrast depth map and generate the 3D-SDP. Instead of just incorporating depth maps into the saliency detection model [41, 42], the proposed method of extracting depth cue has two advantages: (1) the proposed depth measurements produce more informative saliency cues. Useful depth measurements tailored for saliency detection task aim at highlighting salient objects and meanwhile eliminating the background instead of determining the accurate depth value for each pixel or super-pixel. Besides, the proposed 3D-SDP improves the foreground enhancement and background suppression in perspective of rendering the potential salient object center in the image planes and obtains a gradient-like distribution in depth. (2) As [41, 42], the depth saliency is computed based on the depth map. The performance

Table 1 Comparison of average time taken for different saliency detection methods

Methods	LFS	WSC	DLIF	MA	Our approach
Runtime (seconds)	3.8	8.5	0.9	4.2	3.6

may partially depend on an accurate depth map estimation, while the generation of an accurate depth map is time-consuming. Besides, the color saliency map achieves a better result due to the selection of background regions and the proposed 3D-SDP.

To validate the effectiveness of the proposed method, we perform an ablative analysis of our system, by comparing to the following baselines: (1) contrast-based depth saliency, (2) contrast-based depth saliency combined with the 3D-SDP, (3) color saliency, (4) background-based color saliency and (5) background-based color saliency combined with object-biased prior.

Figure 8 shows quantitative comparisons of different baselines. It seems remarkable that the depth saliency maps achieve a better performance (the curves or bars corresponding to the annotation ‘Depth+3D-SDP’) than the color saliency maps (the curves or bars corresponding to the annotation ‘Color+bg+ob’), second only to the fusion of the depth and color saliency. Note that, we just combine the depth and color saliency with a linear method in order to exclude the benefits obtained from various optimization approaches or fusion strategies. Each cue has its own uniqueness to improve the saliency detection performance, even simple combination of depth saliency and color saliency outperforms other methods. Furthermore, experimental results demonstrate that the performance is significantly improved with the proposed 3D-SDP, which strongly validates the effectiveness of the proposed 3D-SDP. Fig. 4 shows some examples where the object does not locate at the image center and results generated from the object-biased prior. It can be seen that our algorithm renders more accurate object center, and therefore better refines the saliency detection results.

However, there exists some limitations of our approach. As mentioned in [20], the salient foreground objects in light fields generally appear ’bigger’ than those in the image-based benchmarks due to the narrow Field-of-View of the Lytro camera. Thus, when object region is small enough, the object center may deviate from its true position. Besides, the proposed object-biased prior would probably degenerate into the center-biased prior when salient objects cannot be distinguished in terms of depth.

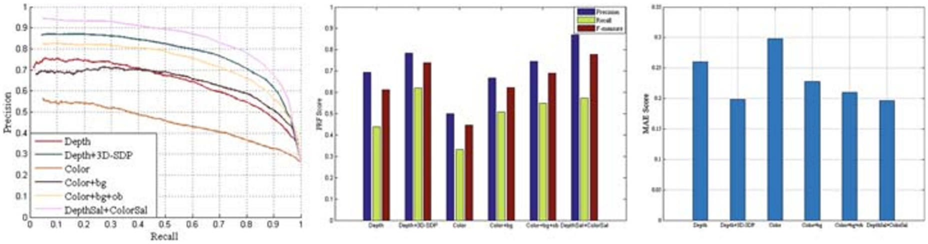


Fig. 8 Quantitative comparisons of different saliency cues in LFSD. **Left**: precision-recall curves from different light field properties; **Middle**: precision, recall, and F-measure for adaptive thresholds. **Right**: MAE scores

6 Conclusion

In this paper, we propose a novel saliency detection model by exploiting the depth cue embedded in a light field. Instead of using the depth map as input or initialization directly, we define the RDFD by computing the degree of defocus for each super-pixel over the light field focal stack. Compared to off-the-shelf approaches relying on accurate depth estimation, the proposed RDFD produces more informative and robust saliency cues mainly in two respects: (1) the regional depth contrast map can be computed by measuring the pair-wise distance between the RDFDs of two super-pixels, (2) the 3D-SDP, including the gradient-like distribution in depth and the object-biased prior in the 2D image plane, can be estimated using RDFDs to further improve the depth saliency map. The RDFD is proved to be an efficient depth measurement tailored for saliency detection task and capable of highlighting salient objects and meanwhile eliminating the background, instead of determining the accurate depth for each pixel. Also, due to the difference operation in computing the RDFD, such measurements eliminate the limitation that the dark channel prior fails when the focused clear image does not contain dark pixels. Experimental results demonstrate that our approach outperforms state-of-the-art methods, and the proposed depth measurement contributes to a significant improvement.

Funding Information This work is supported by NSFC under Grant 61801396 and 61531014, and by the Fundamental Research Funds for the Central Universities under Grant 3102018zy030.

References

1. Achanta R, Estrada F, Wils P, Süsstrunk S (2008) Salient region detection and segmentation. In: ICVS. Springer, pp 66–75
2. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: CVPR, pp 1597–1604
3. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) Slic superpixels compared to state-of-the-art superpixel methods. TPAMI 34(11):2274–2282
4. Alexe B, Deselaers T, Ferrari V (2012) Measuring the objectness of image windows. TPAMI 34(11):2189–2202
5. Borji A, Cheng MM, Jiang H, Li J (2015) Salient object detection: a benchmark. TIP 24(12):5706–5722
6. Bruce N, Tsotsos J (2006) Saliency based on information maximization. In: NIPS, pp 155–162
7. Cheng MM, Mitra NJ, Huang X, Torr PH, Hu SM (2015) Global contrast based salient region detection. TPAMI 37(3):569–582
8. Ciptadi A, Hermans T, Rehg JM (2013) An in depth view of saliency. In: BMVC, pp 112.1–112.11
9. Ding Y, Xiao J, Yu J (2011) Importance filtering for image retargeting. In: CVPR, pp 89–96
10. Donoser M, Urschler M, Hirzer M, Bischof H (2009) Saliency driven total variation segmentation. In: ICCV, pp 817–824
11. Duan L, Wu C, Miao J, Qing L, Fu Y (2011) Visual saliency detection by spatially weighted dissimilarity. In: CVPR, pp 473–480
12. He K, Sun J, Tang X (2011) Single image haze removal using dark channel prior. TPAMI 33(12):2341–2353
13. Itti L (2004) Automatic foveation for video compression using a neurobiological model of visual attention. TIP 13(10):1304–1318
14. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. TPAMI 20(11):1254–1259
15. Jiang B, Zhang L, Lu H, Yang C, Yang MH (2013) Saliency detection via absorbing markov chain. In: ICCV, pp 1665–1672

16. Jiang H, Wang J, Yuan Z, Liu T, Zheng N, Li S (2011) Automatic salient object segmentation based on context and shape prior. In: BMVC, pp 110.1–110.12
17. Jiang P, Ling H, Yu J, Peng J (2013) Salient region detection by ufo: Uniqueness, focusness and objectness. In: ICCV, pp 1976–1983
18. Lang C, Nguyen TV, Katti H, Yadati K, Kankanhalli M, Yan S (2012) Depth matters: influence of depth cues on visual saliency. In: ECCV, pp 101–115
19. Li N, Sun B, Yu J (2015) A weighted sparse coding framework for saliency detection. In: CVPR, pp 5216–5223
20. Li N, Ye J, Ji Y, Ling H, Yu J (2014) Saliency detection on light field. In: CVPR, pp 2806–2813
21. Li X, Lu H, Zhang L, Ruan X, Yang MH (2013) Saliency detection via dense and sparse reconstruction. In: ICCV, pp 2976–2983
22. Liu A, Nie W, Gao Y, Su Y (2016) Multi-modal clique-graph matching for view-based 3d model retrieval. TIP 25(5):2103–2116
23. Liu A, Nie W, Gao Y, Su Y (2018) View-based 3d model retrieval: a benchmark. IEEE Trans Cybern 48(3):916–928
24. Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X, Shum HY (2011) Learning to detect a salient object. TPAMI 33(2):353–367
25. Mai L, Niu Y, Liu F (2013) Saliency aggregation: a data-driven approach. In: CVPR, pp 1131–1138
26. Okutomi M, Kanade T (1993) A multiple-baseline stereo. IEEE Trans Pattern Anal Mach Intell 15(4):353–363
27. Pan J, Sun D, Pfister H, Yang MH (2016) Blind image deblurring using dark channel prior. In: CVPR, pp 1628–1636
28. Peng H, Li B, Xiong W, Hu W, Ji R (2014) Rgbd salient object detection: a benchmark and algorithms. In: ECCV, pp 92–109
29. Pentland AP (1987) A new sense for depth of fiel. TPAMI (4), 523–531
30. Qin Y, Lu H, Xu Y, Wang H (2015) Saliency detection via cellular automata. In: CVPR, pp 110–119
31. Ren J, Gong X, Yu L, Zhou W, Yang MY (2015) Exploiting global priors for rgb-d saliency detection. In: CVPR Workshops, pp 25–32
32. Rutishauser U, Walther D, Koch C, Perona P (2004) Is bottom-up attention useful for object recognition? In: CVPR, vol 2, pp II–II
33. Shen X, Wu Y (2012) A unified approach to salient object detection via low rank matrix recovery. In: CVPR, pp 853–860
34. Sheng H, Feng W, Zhang S (2016) Cellular automata based on occlusion relationship for saliency detection. In: KSEM, pp 28–39
35. Sheng H, Zhang S, Liu X, Xiong Z (2016) Relative location for light field saliency detection. In: ICASSP, pp 1631–1635
36. Wang W, Wang Y, Huang Q, Gao W (2010) Measuring visual saliency by site entropy rate. In: CVPR, pp 2368–2375
37. Xue T, Owens A, Scharstein D, Goesele M, Szeliski R (2019) Multi-frame stereo matching with edges, planes and superpixels. Image Vis Comput 103771:91
38. Yan Q, Xu L, Shi J (2013) Hierarchical saliency detection. In: CVPR, pp 1155–1162
39. Yang C, Zhang L, Lu H, Ruan X, Yang MH (2013) Saliency detection via graph-based manifold ranking. In: CVPR, pp 3166–3173
40. Yang X, Qian X, Xue Y (2015) Scalable mobile image retrieval by exploring contextual saliency. TIP 24(6):1709–1721
41. Zhang J, Wang M, Gao J, Wang Y, Zhang X, Wu X (2015) Saliency detection with a deeper investigation of light field. In: IJCAI, pp 2212–2218
42. Zhang J, Wang M, Lin L, Yang X, Gao J, Rui Y (2017) Saliency detection on light field: a multi-cue approach. TOMM 13(3):32:1–32:22
43. Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. In: CVPR, pp 2814–2821
44. Zitnick C, Kang S, Uyttendaele M, Winder S, Szeliski R (2004) High-quality video view interpolation using a layered representation. In: SIGGRAPH, pp 600–608

Xue Wang is now an assistant professor in School of Computer Science, Northwestern Polytechnical University. She received B.S. in 2007 and Ph.D. in 2017 both from Northwestern Polytechnical University. From 2012 to 2014, she studied in the GRASP lab at the University of Pennsylvania as a visiting joint Ph.D. student financed by China Scholarship Council. Her research interests include computer vision, computational photography and machine learning. She currently focuses building machines that understand the social signals and events that light field videos portray.

Yingying Dong received the B.E. and Master degrees in the School of Computer Science, Northwestern Polytechnical University, in 2016 and 2019. Her research interests focused on visual saliency detection, light field computing and processing during the Master's program.

Qi Zhang received the B.E. degree in Electronic and Information Engineering from Xi’an University of Architecture and Technology in 2013, and Master degree in Electrical Engineering from Northwestern Polytechnical University in 2015. He is now a Ph.D. candidate at School of Computer Science, Northwestern Polytechnical University. His research interests include computational photography, light field imaging and processing, multi-view geometry and applications.

Qing Wang is now a Professor in the School of Computer Science, Northwestern Polytechnical University. He received the B.S. degree from Peking University in 1991. He then joined Northwestern Polytechnical University. In 1997 and 2000 he obtained Master and Ph.D. degrees from Northwestern Polytechnical University. In 2006, he was awarded as outstanding talent program of new century by Ministry of Education, China. He is now a Member of IEEE and ACM. He is also a Senior Member of China Computer Federation (CCF). He worked as research assistant and research scientist in the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University from 1999 to 2002. He also worked as a visiting scholar in the School of Information Engineering, University of Sydney, Australia, in 2003 and 2004. In 2009 and 2012, he visited Human Computer Interaction Institute, Carnegie Mellon University, for six months and Department of Computer Science, University of Delaware, for one month. Prof. Wang’s research interests include computer vision and computational photography, such as 3D reconstruction, object detection, tracking and recognition in dynamic environment, light field imaging and processing. He has published more than 100 papers in the international journals and conferences.