Saliency Detection on Light Field: A Multi-Cue Approach

JUN ZHANG and MENG WANG, Hefei University of Technology
LIANG LIN, Sun Yat-Sen University
XUN YANG and JUN GAO, Hefei University of Technology
YONG RUI, Lenovo

Saliency detection has recently received increasing research interest on using high-dimensional datasets beyond two-dimensional images. Despite the many available capturing devices and algorithms, there still exists a wide spectrum of challenges that need to be addressed to achieve accurate saliency detection. Inspired by the success of the light-field technique, in this article, we propose a new computational scheme to detect salient regions by integrating multiple visual cues from light-field images. First, saliency prior maps are generated from several light-field features based on superpixel-level intra-cue distinctiveness, such as color, depth, and flow inherited from different focal planes and multiple viewpoints. Then, we introduce the location prior to enhance the saliency maps. These maps will finally be merged into a single map using a random-search-based weighting strategy. Besides, we refine the object details by employing a two-stage saliency refinement to obtain the final saliency map.

In addition, we present a more challenging benchmark dataset for light-field saliency analysis, named *HFUT-Lytro*, which consists of 255 light fields with a range from 53 to 64 images generated from each light-field image, therein spanning multiple occurrences of saliency detection challenges such as occlusions, cluttered background, and appearance changes. Experimental results show that our approach can achieve 0.6–6.7% relative improvements over state-of-the-art methods in terms of the *F-measure* and *Precision* metrics, which demonstrates the effectiveness of the proposed approach.

CCS Concepts: • **Computing methodologies → Interest point and salient region detections**; **Computer vision tasks**; **Image representations**; *Computational photography*; Object detection;

Additional Key Words and Phrases: Light field, multi-cue, saliency detection

ACM Reference Format:
Jun Zhang, Meng Wang, Liang Lin, Xun Yang, Jun Gao, and Yong Rui. 2017. Saliency detection on light field: A multi-cue approach. ACM Trans. Multimedia Comput. Commun. Appl. 13, 3, Article 32 (July 2017), 22 pages.
DOI: http://dx.doi.org/10.1145/3107956

1. INTRODUCTION

Saliency detection aims at identifying salient regions or objects that visually stand out from their neighbors. It is a popular area of study, straddling multiple disciplines from

This work is supported by the Natural Science Foundation of China (61403116 and 61432019), the China Postdoctoral Science Foundation (2014M560507), and the Fundamental Research Funds for the Central Universities.

Authors' addresses: J. Zhang, M. Wang (corresponding author), X. Yang, and J. Gao, School of Computer Science and Information Engineering, Hefei University of Technology, 485 Danxia Rd, Hefei 230601, China; emails: {zhangjun1126, eric.mengwang, hfutyangxun}@gmail.com, gaojun@hfut.edu.cn; L. Lin, School of Data and Computer Science, Sun Yat-Sen University (East Campus), Higher Education Mega Center, Guangzhou, China; email: linliang@ieee.org; Y. Rui, Lenovo Group Limited, No. 6 Shang Di West Road, Haidian, Beijing 100085, China; email: yongrui@lenovo.com.

DOI: http://dx.doi.org/10.1145/3107956

cognitive neuroscience to computer vision. In recent years, visual saliency studies have facilitated a growing range of applications such as tracking [6], person re-identification [76], and object detection [39].

Most existing works generally fall into two categories. The first one is based on two-dimensional (2D) visual features, including low-level features such as color, intensity, orientation [69, 74], Gestalt cues [37], and high-level semantic descriptors [13, 68, 77]. Some popular feature learning techniques, such as convolutional neural networks (CNNs) [27, 35, 75], sparse coding [32], and hierarchical neuromorphic networks [61], are also explored to further improve the representation ability of those 2D features for saliency analysis. Although most 2D saliency analysis methods have shown promising results on some widely used 2D datasets [10, 20, 21, 36, 51], they hold some idealistic assumptions, for example, the saliency regions should be occlusion free and should have a different color from their neighborhood/background, the background should be relatively simple and smooth, and so on, which may be not suitable for real-world applications. The other one is three-dimensional (3D) saliency analysis methods [24, 40, 47, 52], which aim to improve the performance by exploiting both depth information extracted from RGB-D (or Kinect[1]) cameras and other 2D features. Along with these methods, some 3D saliency datasets [24, 40, 62] are presented to evaluate the effectiveness of 3D saliency detection methods. However, most existing works may be less effective on cluttered background due to rough depth estimations or when salient objects are situated at distant locations.

Recently, four-dimensional (4D) saliency analysis methods [29, 30, 73] have emerged since the introduction of the Lytro camera,[2] which explore the *light field* [25] for saliency detection. Differing from a regular camera, a Lytro light-field camera can capture a light field towards the scene in a single shot. The light field can be then used to synthesize a focal stack (a stack of images focusing at different depths) and further an all-in-focus image (every pixel is mostly in focus) deriving from the focal stack. Li et al. [30] presented the first and only light-field saliency analysis dataset (*LFSD*) and developed the first light-field saliency detection scheme (LFS) in which they employed the focusness and objectness cues based on the refocusing capability of the light field. They further proposed a weighted sparse coding framework (WSC) in [29] to learn a saliency/non-saliency dictionary, which can effectively handle heterogeneous types of input data, including the light-field data. In [73], Zhang et al. explicitly incorporated depth contrast to complement the disadvantage of color and employed focusness-based background priors to improve the performance of saliency detection. We denote this method as DILF. Although these methods have demonstrated promising performances by leveraging light-field data, accurate saliency detection in real-world scenarios still remains a challenge due to intrinsic and extrinsic factors such as illumination, viewpoint changes, occlusion, and so on. Thus it is necessary to build a more challenging light-field dataset and develop a more effective framework for saliency detection to benefit from the light-field data.

In this article, we propose a new saliency detection scheme that exploits the light-field cues. The architecture overview of our approach is shown in Figure 1. It should be mentioned that the light-field data can be converted into various 2D images (e.g., focal slices, multiple viewpoints, depth maps, and all-in-focus images) [45]. Therefore, we can obtain a list of saliency cues from these light-field images, including the *color* cue from the all-in-focus image, the *depth* cue from the depth map, and the *flow* cues from the focal stack and multiple viewpoints. Moreover, inspired by the influence of center bias on gaze behavior [55, 58] and spatial similarity between scanpaths [17],

[1] www.xbox.com.

[2] http://www.lytro.com/.

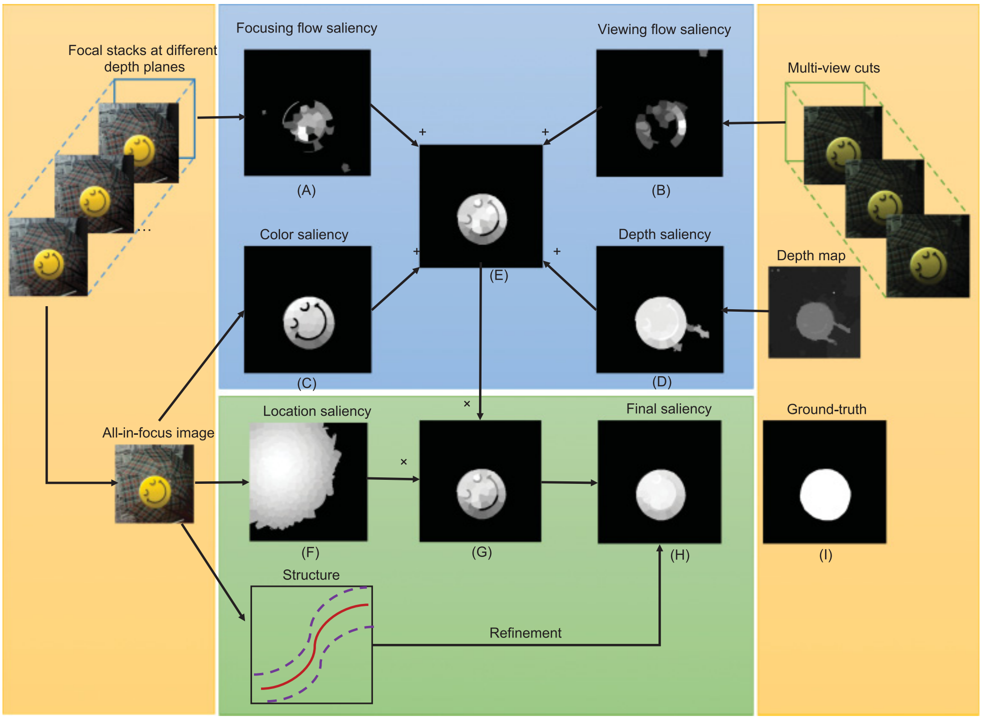

Fig. 1. An illustration of our multi-cue light-field saliency detection model. First, we generate the all-in-focus image, depth map, focal stacks, and multi-view cuts (images) from the corresponding light field captured by a Lytro camera. Light-field flow is estimated from focal slices and multiple viewpoints as a complement to the other information. We measure saliency by computing weighted local contrast, and the prior saliency map is refined by employing a structure cue to obtain the final saliency map.

we also incorporate the *location* prior as a multiplicative weighting factor to enhance the saliency maps. Then we integrate the saliency values obtained from multiple cues using a random-search-based strategy, which can efficiently exploit the complementary of these cues. Finally, we explore a structure-preserving two-stage refinement strategy to refine the saliency map through graph regularization, which encourages adjacent superpixels to take similar saliency values and effectively highlights salient regions. In real-world scenes, these light-field cues are able to complementarily describe the image data from different perspectives.

Specifically, motivated by the computation of *optical flow* [9], we compute the *light-field flow* of light-field data by detecting focus and perspective variations among multiple focal images at different depth planes and multiple viewpoints of the same scene, respectively. The light-field flow can be derived from two components: focusing flow and viewing flow. They can intrinsically encode depth information between the current image and neighboring images. The focusing flow describes depth boundaries between focal slices, and the viewing flow infers depth values from a geometry.

In addition, to better evaluate the existing light-field saliency algorithms, we build a new light-field dataset in this work, named *HFUT-Lytro*, with more realistic and less restrictive image conditions. It is much larger and more challenging than the widely used light-field saliency dataset *LFSD* [30]. It contains 255 challenging scenes, with an average 56 images generated from each light-field scene. These light-field data are captured by a Lytro camera in both indoor and outdoor environments and have a large variance in their appearance and structures. Figure 3 shows some examples from our dataset.

In summary, this article makes the following contributions. (1) We develop a new light-field saliency detection scheme that can effectively and efficiently aggregate color, depth, flow, location, and structure cues from light-field images by using a random search-based weighting strategy. A two-stage refinement framework is also introduced to produce more accurate and structure-preserved results. (2) We propose to estimate the light-field flow from focal and viewpoint sequences to capture depth discontinuities or depth contrast. To the best of our knowledge, we are the first to explore the light-field flow for saliency detection. (3) We present a new light-field saliency detection database, which is the largest and most challenging public dataset for light-field saliency analysis to date. (4) We provide an extensive experimental evaluation on the proposed dataset and the existing LFSD dataset [30], which clearly demonstrates the effectiveness of the proposed saliency detection scheme.

2. RELATED WORK

Saliency detection has been studied extensively in the computer vision community, and readers may refer to [7, 8, 11] for high-quality surveys. Significant progress has also been achieved in light-field saliency detection in recent years. In this section, we only focus on reviewing the most related works.

2.1. Light-field Cameras and Datasets

A light-field imaging system captures not only the projections in terms of intensities but also the directions of incoming light projecting onto an image sensor. Adelson and Bergen [3] presented the plenoptic function to describe the light-field information. We can tell that every 3D point (x, y, z) stores an intensity value, direction, and angle (θ, ϕ), wavelength λ, and time t that a human observer could potentially make at a given moment.[3] Subsequently, Levoy and Hanrahan [25] proposed the two-plane parameterization of the plenoptic function such that each ray is encoded by two parallel planes to represent spatial and angular information. The additional light directions allow the image to be re-focused [45] and the depth information of a scene to be estimated [19, 57, 63]. Therefore, light-field (also known as plenoptic) cameras measure both color and geometric information relative to conventional digital cameras and can operate under some challenging conditions, for example, in bright sunlight.

With the development of modern camera hardware, it has become possible to capture light fields in various ways, including using large camera arrays [67], coded aperture [33], and camera architectures by inserting either a mask [60] or an array of microlenses [44] in front of the photosensor. A review of light-field acquisition devices can be found in [65]. Compared to camera arrays, which are expensive and not very practical, a handheld light-field camera using a microlens array [38, 45], such as Lytro, avoids synchronization and calibration.

Depending on the applications of light fields, current public light-field datasets can be roughly divided into three categories: reconstruction, recognition, and detection. Table I illustrates the main light-field datasets. More information can be found in [64]. At present, LFSD [30] is the only light-field dataset for saliency detection. However, this dataset is not challenging enough, because images are fairly well constrained in terms of illumination, camera location, and so on. A typical scene in the dataset contains only a single centered salient object without occlusions and has very limited clutter, while, in real-world cases, the scene may undergo substantially more complex changes. For instance, the scene may consist of multiple salient objects (e.g., a parking

[3] In this article, we focus on static facts and thus ignore the temporal dimension.

Table I. Overview of Light Field Datasets

Dataset	Number of light fields	Light-field cameras	Applications
The (New) Stanford Light Field Archive[4]	22	A camera array, a gantry and a microscope	Reconstruction (e.g., [59])
Synthetic Light Field Archive[5]	30	POV-Ray	Reconstruction and 3D display (e.g. [41, 66])
Lytro first generation dataset [43]	30	Lytro	Reconstruction (e.g., [43])
LIMU[6]	25	A ProFUSION25 light-field camera	Object detection (e.g., [56])
Kim et al. [23]	5	A Canon EOS 5D Mark II camera	Scene reconstruction (e.g., [23])
GUC-LiFFAD [50]	80	Lyrto	Face recognition (e.g., [50])
LCAV-31 [18]	31	Lytro	N/A
TOLF [70]	18	A camera array	Object recognition and segmentation (e.g., [70, 71])
EPFL Light-Field Image Dataset [54]	118	Lyrto Illum	N/A
HCI [64]	13	Blender, Nikon D800 camera and a gantry	Segmentation (e.g., [42])
LFSD [30]	100	Lytro	Saliency detection (e.g., [29, 30, 73])
HFUT-Lytro	255	Lytro	Saliency detection

lot) or brightness variations (e.g., the reflective highlight on an apple). In this work, we construct a new challenging dataset to facilitate the research and evaluation of visual saliency models using a lenslet light-field camera.

In contrast to the LFSD dataset [30], our proposed dataset has a larger scale and is more challenging, with the real-life scenarios at various distances, sensor noises, unconstrained handheld camera motions, lighting conditions, and so on. Moreover, this dataset also offers multi-view images with the corresponding camera calibration file, which are not provided in the LFSD dataset.

2.2. Saliency Detection with Multiple Visual Cues

It has been acknowledged that multiple cues can be explored to benefit saliency detection. For example, Zhao and Koch [74] combined different feature channels (color, intensity, orientation, and face) across multiple spatial scales in a nonlinear AdaBoost framework on three 2D datasets. They found that the performance of the saliency model can be consistently improved by the nonlinear feature integration and a center bias model. Xu et al. [68] used a linear Support Vector Machine (SVM) classifier to combine pixel-, object-, and semantic-level attributes and demonstrated the importance of the object- and semantic-level information. Another work by Ma and Hang [40] also used a linear SVM classifier to integrate low-, mid-, and high-level features to predict saliency maps on the NCTU-3DFixation dataset. Ren et al. [52] proposed to integrate depth, orientation, and background priors with the region contrast to produce saliency maps in a linear manner. Recent progress in CNNs have boosted the performance of saliency detection by automatically learning hierarchical features. For example, Li and Yu [27] concatenated CNN features at multiple scales for saliency prediction. Liu et al. [35] developed a CNN architecture with multiple resolutions to simultaneously learn

[4] http://lightfield.stanford.edu/lfs.html.
[5] http://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php.
[6] http://limu.ait.kyushu-u.ac.jp/dataset/en/lightfield_dataset.htm.

early features, bottom-up saliency, and top-down factors and integrated them into the logistic regression layer to predict eye fixations.

Compared with previous studies, we acquire and integrate visual cues from a single light-field camera. The integrated visual information allows the visual saliency to be predicted in a global plenoptic space.

3. OUR APPROACH

In this section, we present an effective and efficient saliency detection scheme using multiple cues from a Lytro light-field camera. Given a light-field sample, consisting of an all-in-focus image I_{af}, a depth image I_{depth}, a focal stack, and multi-view cuts (images), the task is to obtain a saliency map with regard to the all-in-focus image. We aim to explore multiple saliency cues deriving from the light-field data to detect the saliency in the all-in-focus image.

We begin with segmenting the all-in-focus image I_{af} into a set of non-overlapping regions/superpixels $\{sp_i\}_{i=1}^{M}$ by the widely adopted simple linear iterative clustering (SLIC) algorithm [2], where $M = h \times w/N$ is the superpixel number, N is the number of pixels within each superpixel, and h and w are the height and width of the current image, respectively. Here, the superpixels usually have a more regular and compact shape with better boundary adherence than the uniformly segmented regions. In the following sections, our major goal is to compute the saliency map $S(sp_i)$ of the superpixel $sp_i, i = 1, \ldots, M$. In this work, we use k to index four saliency cues: color cue, depth cue, and two flow cues (focusing flow and viewing flow).

Our approach mainly consists of the following steps.

(1) First, for each cue, we estimate the superpixel-level feature distinctiveness $d^k(sp_i, sp_j)$ between superpixel sp_i and superpixel sp_j over the corresponding light-field image plane, which is described in detail in Section 3.1. Here, the feature distinctiveness is a pairwise distance that measures the feature difference between two superpixels/regions in the corresponding feature (cue) space.

(2) Second, we incorporate the location prior [53, 73] into our scheme by computing two location cues: *implicit location* $\ell_{im}(sp_i, sp_j)$ in Equation (3) and *explicit location* $\ell_{ex}(sp_i)$ in Equation (4), where $\ell_{im}(sp_i, sp_j)$ measures the spatial similarity between two superpixels and $\ell_{ex}(sp_i)$ measures the center bias by computing the Gaussian distance between the centers of the superpixel sp_i and the all-in-focus image. A detailed procedure is described in Section 3.2.

(3) Then, we can compute the initial saliency map of each superpixel over all the feature (cue) spaces by

$$S^*(sp_i) = \sum_k \sum_{j=1}^{M} \alpha_k \ell_{ex}(sp_i) \ell_{im}(sp_i, sp_j)\, d^k(sp_i, sp_j),\ j \neq i, \tag{1}$$

where $\alpha_k > 0$ is the weight assigned to the cue k, $\sum_k \alpha_k = 1$.

The basic idea behind Equation (1) is that the superpixel that differs from its neighboring regions should correspond to high saliency values. In Equation (1), the implicit location cue $\ell_{im}(sp_i, sp_j)$ is used to weight the feature distinctiveness $d^k(sp_i, sp_j)$ by considering that closer regions should have a higher voting weight. The explicit location cue $\ell_{ex}(sp_i)$ also functions as a weight factor by following an assumption that a region is more likely to be salient if it is close to the center of the image.

The cue weight α_k is determined by a random search of T trials that maximizes the F-measure defined in Equation (7). In comparison with grid search, random

```
ALGORITHM 1: The Proposed Light-field Saliency Detection Approach
Input:
A light-field sample: all-in-focus image, depth map, a stack of focal slices, and multiple viewpoints;
$\sigma_\mu$, $\sigma_c$, $\sigma_s$, $\sigma_a$: radius parameters of Gaussian function;
$\gamma$, $\lambda$: regularization parameters;
$\{\alpha_k\}$: light-field cue weights; /* Note that the cue weight $\alpha_k$ is optimized by random
    search strategy on validation set that maximizes the F-measure defined in
    Equation (7).                                                                    */
$\beta$: the trade-off between precision and recall;
Output: The final saliency map $S_{opt}$;
 1: Over-segment the all-in-focus image to obtain $M$ superpixels $\{sp_i\}_{i=1}^{M}$;
 2: Estimate the feature distinctiveness between two superpixels for each light-field cue by Equation (2) in Section 3.1;
 3: Compute the implicit and explicit location priors by Equation (3) and Equation (4) in Section 3.2;
 4: Obtain the initial saliency map by Equation (1) and normalize it to a range of $[0, 1]$;
 5: Refine the initial saliency map by the two-stage refinement strategy in Section 3.3;
Return: $S_{opt}$;
```

search over the same domain is able to achieve a comparable or even better performance at a lower computational cost [5].

(4) Next, we normalize the initial saliency map to a range of $[0, 1]$ by $\mathcal{N}(\mathbf{x}) = \frac{\mathbf{x}-\min(\mathbf{x})}{\max(\mathbf{x})-\min(\mathbf{x})}$, where $\mathbf{x}$ represents the vector of all superpixels' saliency values. The normalized initial saliency map is denoted by $S(sp_i)$. We consider the region in the whole saliency map that has a larger value than a global threshold to belong to the salient region based on Otsu's method [46].

(5) Finally, we apply a two-stage structure-preserving refinement strategy to produce the final saliency map, which is described in detail in Section 3.3. To make our approach clearer, we summarize the proposed light-field saliency detection scheme in Algorithm 1.

3.1. Superpixel-level Intra-cue Distinctiveness

In this subsection, we will introduce how to derive the superpixel-level distinctiveness of the color, depth, and flow cues from the light-field data. For a cue k, the superpixel-level distinctiveness is measured by

$$d^k(sp_i, sp_j) = \left\| m^k(sp_i) - m^k(sp_j) \right\|_2, j = 1, \ldots, M \text{ and } j \neq i, \tag{2}$$

where sp_i and sp_j denote the ith and jth superpixels, respectively, in the corresponding light-filed image. $m^k(\cdot)$ denotes the average feature (color, depth, or flow) values of each superpixel for cue k. The color, depth, and flow cues are described, respectively, as follows.

Color. Color is the most widely used cue in saliency detection. It measures how the intensity of incoming light varies with wavelength. In this work, the color measurement is derived from the all-in-focus image I_{af} in which each pixel is encoded by RGB values. We first transform the RGB image I_{af} into the Lab color space. Then we estimate the color distinctiveness to characterize common color changes over each channel. The final color distinctiveness between two superpixels in the all-in-focus image I_{af} is equivalent to the summation of pairwise distances over three color channels, $d^{color}(sp_i, sp_j) = \sum_c d_c^{color}(sp_i, sp_j)$, where c indexes the three color channels of Lab color space.

Depth. For a scene with similar foreground and background colors, color cue becomes less useful. To address this situation, we capture the depth information from the light-field depth image I_{depth}. Each pixel in the depth image describes the distance of the surface of a object from a viewpoint. Based on the observation that the region at a closer depth range tends to be more salient, we compute the depth distinctiveness $d^{depth}(sp_i, sp_j)$ between two superpixels from the light-field depth image I_{depth} using Equation (2).

Flow. The most significant characteristic of a light-field camera is that it can generate a stack of images focusing at different depths and multiple viewpoints towards the scene on a 2D sampling plane in one shot, which allows the estimate of flow field (flow velocities at horizontal and vertical directions) for the light-field data. Besides, salient regions usually are closer to the camera than background regions, and, in general, they will display more apparent motion than background regions, which makes the flow information of light-field data an effective cue for saliency detection. In this work, we apply the computational work of *optical flow* [9] to estimate the *light-field flow* of 4D light-filed data. To the best of our knowledge, we are the first to explore the light-field flow for saliency detection.

For a light-field image sequence $\{I_f\}_{f=1}^{F}$ (a stack of focal slices or multiple viewpoints shown in Figure 1), we describe the light-field flow at the pixel (x, y) by a 3D vector field $(\triangle x, \triangle y, 1)$, which is the displacement vector between two consecutive images (I_f and I_{f+1}). Here, F is the number of frames in the current focal stack or viewpoints and $\triangle x$ and $\triangle y$ are the horizontal and vertical components at the pixel (x, y), respectively. We then compute the flow displacement $(\triangle x, \triangle y)$ by the optimization algorithm in [9]. Then we can estimate the light-field flow map as the average value of the square root of flow displacements over all the focal slices or the viewpoints, $\frac{1}{F}\sum_{f=1}^{F}\sqrt{\triangle x^2 + \triangle y^2}$. We term the light-field flow map derived from the stack of focal slices as *focusing flow* and the light-field flow map derived from all the viewpoints as *viewing flow*. These light-field flow maps will make a good complement to other saliency cues for better saliency detection. After obtaining these flow maps, we can calculate the focusing flow distinctiveness $d^{focusFlow}(sp_i, sp_j)$ and the viewing flow distinctiveness $d^{viewFlow}(sp_i, sp_j)$ between two superpixels using Equation (2), respectively.

3.2. Implicit and Explicit Location Measures

Although the above cues are able to detect salient regions, they mainly rely on the contrast information, which may be less effective when handling low-contrast images. Several studies have suggested that the relevance between pixels/regions is increased when their spatial distance is decreased [53], and people often focus their eyes onto the center of an image [21, 58]. In this work, we aim to enhance the ability of saliency detection by introducing two location-based saliency priors: *implicit location* and *explicit location* as follows.

Implicit Location. We compute the spatial similarity $\ell_{im}(sp_i, sp_j)$ between the centers of two superpixels sp_i and sp_j by

$$\ell_{im}(sp_i, sp_j) = \exp\left(-\frac{1}{2\sigma_{\mathbf{u}}^2}\|\mathbf{u}(sp_i) - \mathbf{u}(sp_j)\|_2^2\right), \tag{3}$$

where $\mathbf{u}(\cdot) = (\overline{x}/h, \overline{y}/w)^T$ denotes the normalized center coordinate of the given superpixel, where $(\overline{x}, \overline{y})$ denotes the average center coordinates that is computed by averaging the coordinates of all the pixel within the superpixel, h and w denote the height and weight of the all-in-focus image, and $\sigma_{\mathbf{u}}$ is a radius parameter.

Explicit Location. We measure the center bias by calculating the distance between the image centroid $\mathbf{u}_c$ and the center coordinate $\mathbf{u}(sp_i)$ of superpixel sp_i. We additionally exploit the background prior to enhance the center-bias salient locations and suppress non-salient regions. For this purpose, we calculate the highest background probability $m^{bg}(sp_i)$ for each superpixel from the corresponding focal slice based on the method in [73], and we use this probability to improve the center-bias modeling. Similarly to the implicit location, we use a Gaussian function to obtain the location weight,

$$\ell_{ex}(sp_i) = \exp\left(-\frac{1}{2\sigma_c^2}\left(1 - m^{bg}(sp_i)\right)^2 \|\mathbf{u}_c - \mathbf{u}(sp_i)\|_2\right), \tag{4}$$

where σ_c is a radius parameter.

The implicit and explicit location priors function as two spatial weight factors in Equation (1) to enhance the performance of saliency analysis. With this enhancement operation, the values of salient regions in images are increased, while the saliency values of non-salient background regions are decreased.

3.3. Saliency Refinement by Exploring the Structure Cue

Although, by integrating light-field saliency cues, some salient parts of regions can be identified, the whole object may not be uniformly highlighted, as shown in Figure 1(G). In addition, the saliency maps usually include fuzzy object boundaries and background noises. The third column in Figure 10(A) shows four saliency detection results produced by the proposed scheme based on the integration of color, depth, flow, and location cues. To alleviate these problems, we further propose to explore the light-field structure cue to refine the saliency values by considering the interrelationship between adjacent elements (e.g., neighboring nodes are more likely to share similar visual properties and saliency values).

We utilize internally homogeneous and boundary-preserving superpixels as basic representation units and refine the normalized initial saliency map $S(sp_i)$ by exploring the salient (foreground) term $S(\cdot)$ and the background (non-salient) term $S_{bg}(\cdot)$ based on the boundary connectivity [78]. Specifically, we optimize $S(sp_i)$ by

$$\begin{aligned}\min_{S_{opt1}(sp_i)} \quad & \sum_{i=1}^{M} S(sp_i)\|S_{opt1}(sp_i) - 1\|^2 + \sum_{i=1}^{M} S_{bg}(sp_i)\|S_{opt1}(sp_i)\|^2 \\ & + \sum_{i,j=1}^{M} w(sp_i, sp_j)\|S_{opt1}(sp_i) - S_{opt1}(sp_j)\|^2.\end{aligned} \tag{5}$$

Here, the first term defines the cost that encourages a superpixel sp_i with a large foreground saliency probability $S(sp_i)$ to take a large saliency value $S_{opt1}(sp_i)$. The second term defines the background cost with a small saliency value $S_{opt1}(sp_i)$. The last term encourages the adjacent superpixels to have similar saliency values, in which

$$w(sp_i, sp_j) = \exp\left(-\frac{1}{2\sigma_s^2}\sum_k d^k(sp_i, sp_j)\right) + \gamma \exp\left(-\frac{1}{2\sigma_a^2} d^{color}(sp_i, sp_j)\right), \tag{6}$$

where the first term measures the similarity across different cues, the second term is the color affinity of every adjacent superpixel pair, γ is a small regularization parameter, and σ_s and σ_a modulate the strengths of the weights.

Equation (5) is a nonlinear least-square regression problem, which can be easily solved to obtain a closed-form solution. While most regions of the salient objects are detected in this refinement stage (R1), some background regions may not be adequately

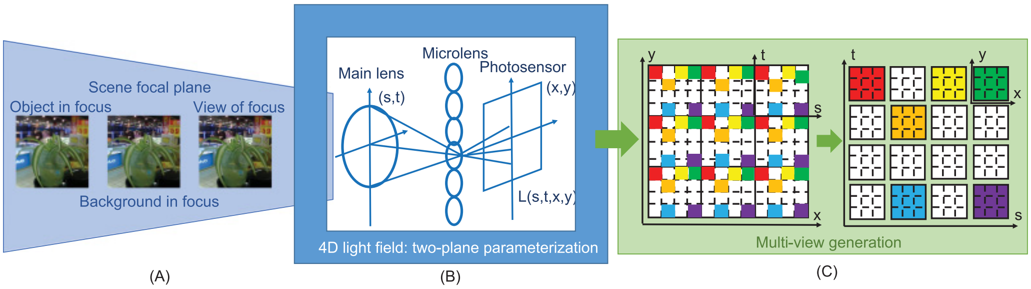

Fig. 2. An illustration of the light-field imaging principle. From left to right: slices at different focusing planes, a schematic illustration of a light-field camera, and multi-view generation.

suppressed (see the fourth column of Figure 10(C)). To improve the performance, in the second stage (R2), we follow the general-purpose undirected graph construction framework [26, 72] by using the region saliency S_{opt1} as the new query to construct a saliency map S_{opt2}. Finally, the optimized saliency map is obtained by the scalar production of the two steps of saliency refinement results, that is, $S_{opt} = S_{opt1} \cdot S_{opt2}$. As a result, the saliency of each superpixel, as measured by the edges to which it belongs, is not only determined by the superpixel itself but also influenced by its associated structural contexts.

4. HFUT-LYTRO DATASET

This section introduces HFUT-Lytro, a light-field dataset collected using a commercially available first-generation Lytro camera. This camera is composed of a main lens, a microlens array, and a photosensor [45], as shown in Figure 2(B). Our main focus is to explore the inherent characteristics of the light-field camera, which can be used to identify salient regions in real-world sequences. To build this dataset, we initially collect 360 light fields from a variety of indoor and outdoor scenes. A total of 255 samples are then chosen manually based on at least one of the following criteria: (i) multiple disconnected salient object regions; (ii) various depths such that salient regions are allowed to be in front of the camera at a distance of 2–3m in real backgrounds; (iii) more illuminants, such as highlight or dark light; (iv) partially occluded salient regions; (v) cluttered background or the similar foreground and background; and (vi) small-scale salient regions.

The light field is the total spatio-angular distribution of light rays passing through the free space and can be parameterized in a two-plane parametrization $L_F(x, y, s, t)$ [25]: the viewpoint plane (s, t) and the image plane (x, y). For each sample in this dataset, multiple images with different regions in focus are obtained from the Lytro camera based on the digital refocusing technique [45]. We show three focal slices corresponding to different depth planes in Figure 2(A). Then, a single all-in-focus image with sharp focus at every pixel can be created by combining these focal images [4], and the depth of each ray of light can be estimated by measuring pixels in the focus images [57]. Moreover, multi-view images are generated by scanning all the spatial locations at any given viewpoint, as shown in Figure 2(C). For example, the view marked in red is generated from all the spatial coordinates (x, y) at the given viewpoint (s, t). In our case, each light-field sample in the proposed dataset contains 49 sub-aperture images, which have a 7×7 angular resolution and 328×328 pixels of spatial resolution.

Six examples from our dataset are shown in Figure 3. Figures 3(A) and (F) illustrate all-in-focus images and depth maps generated by Lytro Desktop 3.0[7], Figures 3(B)–(D)

[7] https://www.lytro.com/desktop.

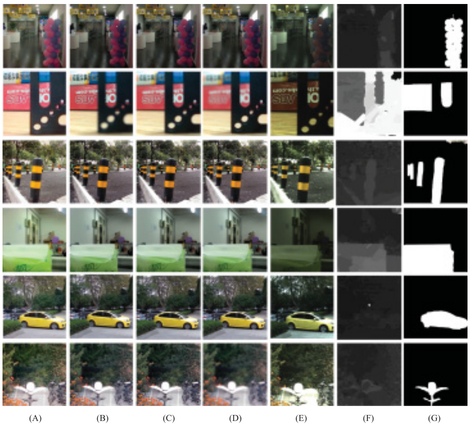

Fig. 3. Examples in the HFUT-Lytro dataset. From (A) to (G): all-in-focus images, focal slices sampled from different depth planes, center viewing images, depth maps, and human-masked ground truths.

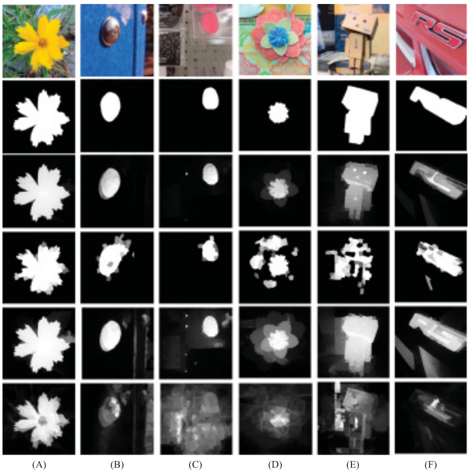

Fig. 4. Saliency detection results of different methods on the LFSD dataset. From top to bottom: all-in-focus images, ground-truth maps, and saliency maps obtained by our approach, WSC [29], DILF [73], and LFS [30].

show three slices focusing at different depth planes generated by the Lfptools.[8] Figure 3(E) gives the center-viewing image from multi-view samples using the decoding method proposed in [15]. We create ground truths by human subjects. Five subjects, 3 males and 2 females, are asked to draw bounding contours around the objects/regions that attract them in the image. It is noticed that some regions marked by the subjects have inconsistencies in terms of being salient ground truths such as the regions where there are multiple objects in the scene. Hence, we set the pixel value to 1 if at least 3 subjects agree that it belongs to a salient region and zero otherwise. Since one should respond uniformly within the whole region [28], finally, one participant uses Adobe Photoshop to segment the salient region maps manually from each image. Figure 3(G) shows the corresponding ground-truth saliency maps.

5. EXPERIMENTS

In this section, we first introduce the datasets and metrics used for the evaluation and the implementation details in Section 5.1. We then compare our approach with state-of-the-art methods on two datasets in Section 5.2. In Section 5.3, we investigate the effects of different light-field cues for saliency detection and verify the effectiveness of multi-cue fusion methods and superpixel segmentation.

5.1. Experimental Settings

5.1.1. Datasets. To evaluate the performance of the proposed saliency detection approach, we perform extensive evaluations on the proposed HFUT-Lytro dataset and the LFSD dataset [30].

HFUT-Lytro. This dataset contains 255 light fields. As previously mentioned, most scenes contain multiple objects appearing at a variety of locations and scales with

[8] https://github.com/nrpatel/lfptools.

complicated background clutter, which makes this dataset more challenging for saliency detection.

LFSD. This dataset consists of 100 light fields of different scenes. It was originally designed for saliency detection captured with a Lytro camera. Most of the scenes contain only one salient object with high contrast and constrained depth ranges.

5.1.2. Evaluation Criteria. We utilize four metrics for quantitative performance evaluations: Precision and Recall (PR) curve, average precision (AP), F-measure, and Mean Absolute Error (MAE). A PR curve is obtained by binarizing the saliency map using different thresholds (ranging from 0 to 255), resulting in a pair of precision and recall values when the binary mask is compared against the ground truth. The curves are averaged over each dataset. The precision score reflects the constancy of the saliency map in predicting the locations within the ground truth. A high precision score shows that most of the predicted locations are true positives. However, high compactness in the predicted locations will result in a low recall, which indicates minimal coverage of the regions enclosed by the salient region's boundary. AP is defined as the area under the PR curve. Following [7], we use the PASCAL evaluation protocol [16] to evaluate this performance.

The F-measure summarizes precision and recall information in a single value. Traditional F1-measure may have limitations in several cases, since recall and precision are evenly weighted [49]. Therefore, we use the general F_β measure to characterize the tradeoff between precision and recall such that

$$F_\beta = \frac{(1+\beta^2)P \times R}{\beta^2 \times P + R}. \tag{7}$$

As suggested in [1, 14], we set $\beta^2 = 0.3$ to weight precision more than recall. P and R are the precision and recall rates obtained by an adaptive threshold that is twice the mean saliency of the image [1], that is, $Th = \frac{2}{h \times w} \sum \sum S_{opt}$. Using this method, we obtain average precision and recall over each dataset. We following the same settings in [31] in our implementation.

MAE provides a better estimate of the dissimilarity between the continuous saliency map S_{opt} and the binary ground-truth GT [48], which is defined as

$$MAE = \frac{1}{h \times w} \sum \sum |S_{opt} - GT|. \tag{8}$$

5.1.3. Implementation Details. In our experiments, each input all-in-focus image is normalized to have zero mean and unit variance. The pixel number N within each superpixel is set to 400. We use the code provided by [34] to compute the light-field flow and set the parameter $\lambda = 0.012$. γ is empirically set to 0.1, and the parameters of the RBF are set as follows: $\sigma_\mu = 2/3$, $\sigma_c = 1$, and $\sigma_a = 1$. σ_s is adaptively determined using the method in [78]. The weight vector α is set to $[0.50, 0.17, 0.23, 0.10]$ and $[0.54, 0.33, 0.13]$ after 100 trials of the random search on the HFUT-Lytro and LFSD datasets, respectively (note that there is no viewing flow information supported by the LFSD dataset).

5.2. Comparison with State-of-the-art Methods

We compare the proposed approach with three recent state-of-the-art methods, LFS [30], WSC [29], and DILF [73], on the LFSD and HFUT-Lytro datasets. We use either the implementations with recommended parameter settings provided by the authors or the results provided by the authors for comparisons.

5.2.1. Quantitative Results. The PR curves of all the methods on the two datasets are shown in Figure 5. The results show that our method achieves a higher precision

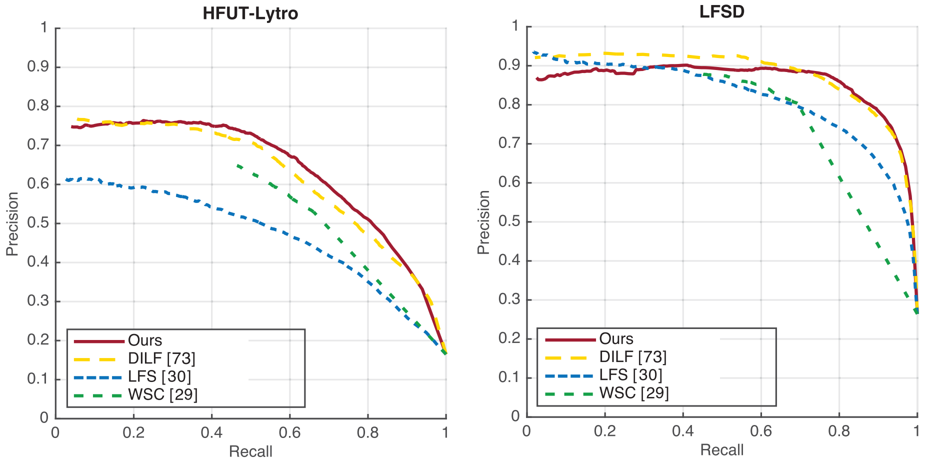

Fig. 5. The PR curves obtained by different methods on the HFUT-Lytro (left) and LFSD (right) datasets.

Table II. The Precision, Recall, F-measure, AP, and MAE Obtained by Different Methods on the HFUT-Lytro and LFSD Datasets (Bold: Best; Underline: Second Best)

Dataset	Metric	Ours	DILF[73]	WSC[29]	LFS[30]
HFUT-Lytro	Precision	**0.5928**	0.5186	0.5254	0.4753
	Recall	0.6726	**0.7147**	0.6673	0.5354
	F-measure	**0.6095**	0.5537	0.5525	0.4880
	AP	**0.6354**	0.6221	0.4743	0.4718
	MAE	**0.1388**	0.1578	0.1454	0.2214
LFSD	Precision	**0.8542**	0.8271	0.8076	0.8115
	Recall	0.7397	**0.7916**	0.6783	0.6083
	F-measure	**0.8247**	0.8186	0.7735	0.7534
	AP	0.8625	**0.8787**	0.6832	0.8161
	MAE	0.1503	**0.1363**	0.1453	0.2072

in almost the entire recall range on the HFUT-Lytro dataset. On the LFSD dataset, DILF outperforms our approach. Furthermore, the results of the precision, recall, F-measure, AP, and MAE metrics are shown in Table II. Our approach performs favorably against the state-of-the-art methods in most cases. More specifically, our approach outperforms all the other methods in terms of the F-measure and precision metrics. On average, our approach obtains improved performances of 5.6% and 0.6% over the best-performing state-of-the-art algorithm (DILF) in terms of F-measure and performs better than the second-best method by 6.7% and 2.7% according to the precision score on the HFUT-Lytro and LFSD datasets, respectively. One explanation for this is that most methods that utilize color and/or depth cues suffer from low precision, since it is sometimes difficult to distinguish salient regions and distractors with similar appearances. Therefore, the distractors will also pop out along with the salient regions, leading to a decreased precision. By combining the results from Figure 5 and Table II, we can see that our approach can yield more significant improvements on the more challenging HFUT-Lytro dataset.

5.2.2. Qualitative Results. For an intuitive comparison, we provide several representative saliency maps generated by our approach and other state-of-the-art methods on the LFSD and HFUT-Lytro datasets in Figure 4 and Figure 6, respectively. Clearly, most of the saliency detection methods can handle images with relatively simple backgrounds

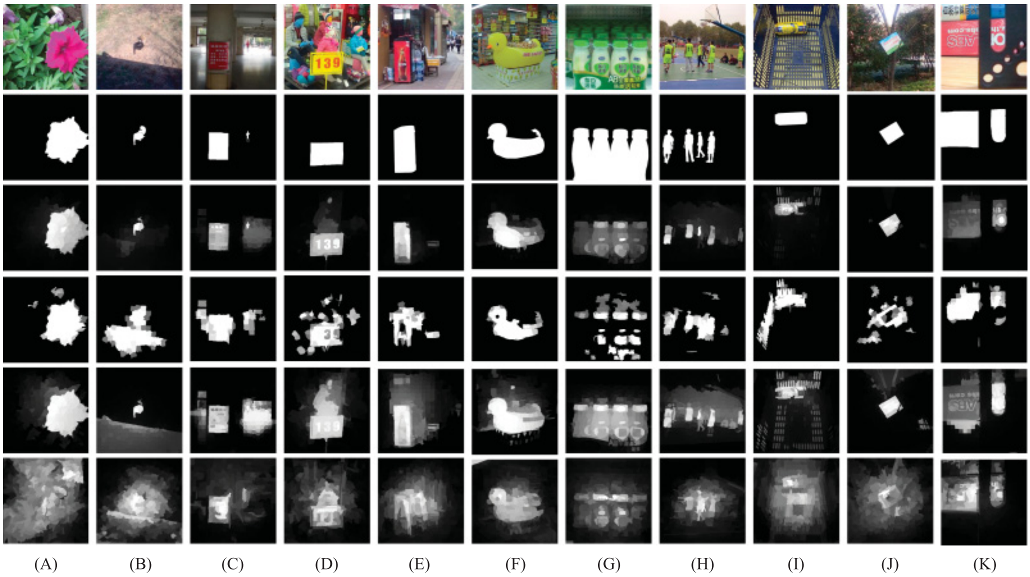

Fig. 6. Saliency detection results of different methods on the HFUT-Lytro dataset. From top to bottom: all-in-focus images, ground-truth maps, and saliency maps obtained by our approach, WSC [29], DILF [73], and LFS [30].

and homogeneous objects, as shown in Figure 4(A). However, our approach is capable of handling various challenging cases. For example, although the contrast between objects and backgrounds is very high in Figure 4(B) and Figure 6(A), some methods cannot accurately highlight the salient regions due to background noise, whereas our approach can effectively address this issue. The proposed approach also performs better than other methods, even when the salient regions are distracted by background illumination and shadow artifacts, as shown in Figure 4(C) and Figures 6(B) and (C). Moreover, in the cases where salient regions are distracted from a cluttered background or have similar appearance with backgrounds (such as Figures 4(D)–(F) and Figures 6(D)–(F)), our approach is able to highlight the salient parts more coherently and provides a better prediction. Moreover, our approach still performs well when the background prior is invalid to a certain extent, for example, multiple objects are present in the same scene (Figures 6(G) and (H)), the salient objects are located at the distant depth levels (Figure 6(I)), are very small (Figures 6(B) and (J)), or are occluded by nearby objects (Figure 6(K)).

5.3. Analysis of the Proposed Approach

In this section, we evaluate our approach from multiple perspectives. In the following experiments, the contribution of each component in our approach is discussed.

5.3.1. On the Contribution of Individual Cues. We first evaluate each individual cue separately. For each cue, we set its weight to one and the weights of other cues to zero using the same scheme described in Section 3. The results are shown in Figure 7 and Figure 8.

We can observe in Figure 8 that the color cue quantitatively performs much better than the other cues in terms of precision, recall, F-measure, and MAE on the HFUT-Lytro dataset. However, the performance of the depth cue is slightly better than that of the color cue on the LFSD dataset in Figure 7. We also noticed that the depth cue is

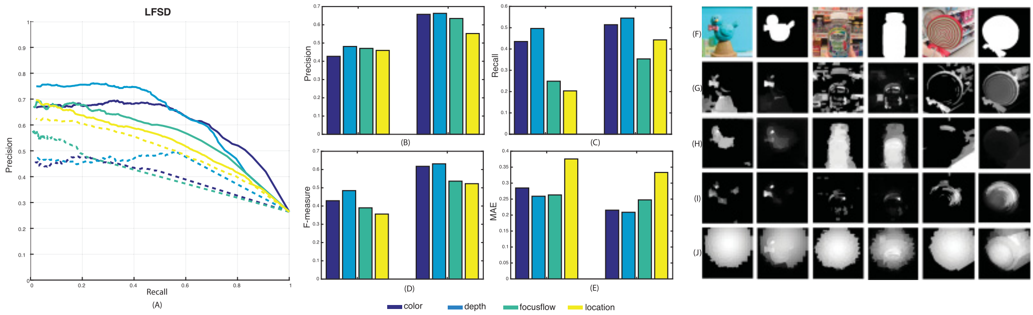

Fig. 7. Quantitative and qualitative comparisons of saliency maps from our framework with individual light-field cues and their refinements on the LFSD dataset. (A) PR curves of individual saliency maps (dashed line) and their refinements (solid line); (B)–(E) quantitative results of individual saliency maps (left) and their refinements (right) for each metric; (F) all-in-focus images and the corresponding ground-truth saliency maps; (G)–(J) color-, depth-, focusing flow-, and position-driven saliency maps and their refined versions with the structure cue.

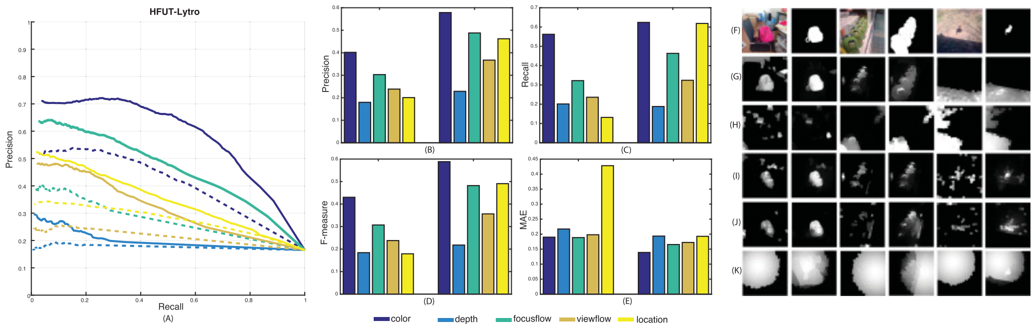

Fig. 8. Quantitative and qualitative comparisons of saliency maps from our framework with individual light-field cues and their refinements on the HFUT-Lytro dataset. (A) PR curves of individual saliency maps (dashed line) and their refinements (solid line); (B)–(E) quantitative results of individual saliency maps (left) and their refinements (right) for each metric; (F) all-in-focus images and the corresponding ground-truth saliency maps; (G)–(K) color-, depth-, focusing flow-, viewing flow-, and position-driven saliency maps and their refined versions with the structure cue.

essential for certain images. Two such examples are shown in the first two examples of Figure 7. In the first sample image, the interesting object has similar colors with background so the color cue cannot provide enough discriminative information. And the second example shows that the complex-textured background makes it hard for the model to locate the salient object without depth information. The depth cue helps to allocate them and, consequently, the accuracy of the saliency model is significantly improved in these cases. However, if the object has a more complex pattern with a large size (i.e., the number of superpixels of the object is more than that of the background), the depth fails to predict the saliency region of the object, as shown in the third example of Figure 7. In this case, color and focusing flow cues can detect object boundaries based on the color variations and depth discontinuity to form a coarse saliency map. Then the refinement is capable of rendering a uniform saliency map while suppressing the background, leading to salient objects being popped out. On the other hand, we observe that some images are especially hard to predict using the depth cue. Two examples are shown in Figure 8. A group of pots in the second example are located at large depth planes. Even though the color cue can recognize the five pots, the depth feature only detects the nearer pot to the observers. A similar phenomenon appears on the

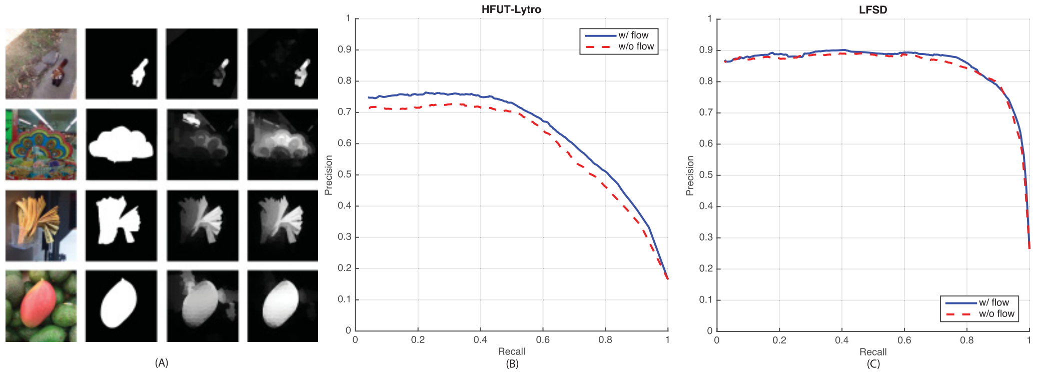

Fig. 9. PR-curve and qualitative comparisons of saliency maps with and without the flow cue. (A) Qualitative comparison of saliency maps. From left to right: all-focus images, ground truth, w/o the flow cue, and w/ the flow cue; (B) HFUT-Lytro dataset; (C) LFSD dataset.

third sample image, in which the animal is too small and located in the further depth plane so the depth cue misidentifies some regions. Besides, the color cue points out the shadow in the scene. We find such phenomena also have connections with the characteristics of the datasets. Salient objects in the LFSD dataset are closer to the camera. This leads to more accurate and useful depth estimation for saliency detection. However, in the HFUT-Lytro dataset, salient objects have wider depth ranges. Thus, depth estimation could be inaccurate, and they could treat closer regions as salient regions. By examining these predicted saliency maps, we can see that the flow features are critical to the specific cases. Moreover, the location saliency maps based on the Gaussian function have a similar appearance but different resolutions and aspect ratios. Note that the aliasing effect is caused by thresholding the saliency maps. We also illustrate the comparison of saliency detection with and without structure-cue-based saliency refinement for individual cue. We find that the refinement leads to a better saliency detection performance in all metrics. It is also clear from visual comparisons that the structure cue is able to highlight the salient object parts more coherently for each other cue and achieves a better prediction, especially on complex scenes with cluttered background.

From the results in Figure 5, Figure 7, and Figure 8, we can see that the cue integration yields better results than using any individual cue alone. Different visual cues provide complementary supporting information to saliency detection.

5.3.2. Evaluation on the Flow Cue. Here, we evaluate the contribution of the flow cue to the accuracy of the overall combination. For this purpose, we simply remove this cue from the full scheme with the other cues left unchanged with equal weights for fair comparison. Hence, the more the performance is decreased, the more important the flow cue is to the overall accuracy. Some saliency maps generated with and without light-field flow cues are shown in Figure 9(A), and the resulting quantitative comparisons are shown in Figures 9(B) and (C).

It can be seen from Figures 9(B) and (C) that the performance of saliency detection significantly degrades after removing the flow cue from the combination, especially on the HFUT-Lytro dataset. This well demonstrates the effectiveness of flow information in the visual saliency measurement. We also observe in Figure 9(A) that the flow cue does not respond against the textured background but can detect the depth boundary. Hence, when all cues are combined, this boundary is assigned a higher probability of being saliency pixels than the neighboring background.

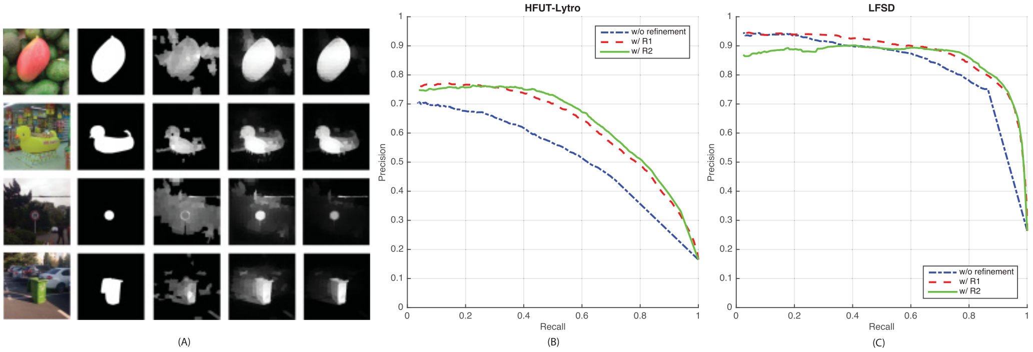

Fig. 10. PR-curve and qualitative comparisons of saliency maps with and without refinement. (A) Qualitative comparison of saliency maps. From left to right: all-focus images, ground truth, w/o refinement, w/ R1, and w/ R2; (B) HFUT-Lytro dataset; (C) LFSD dataset.

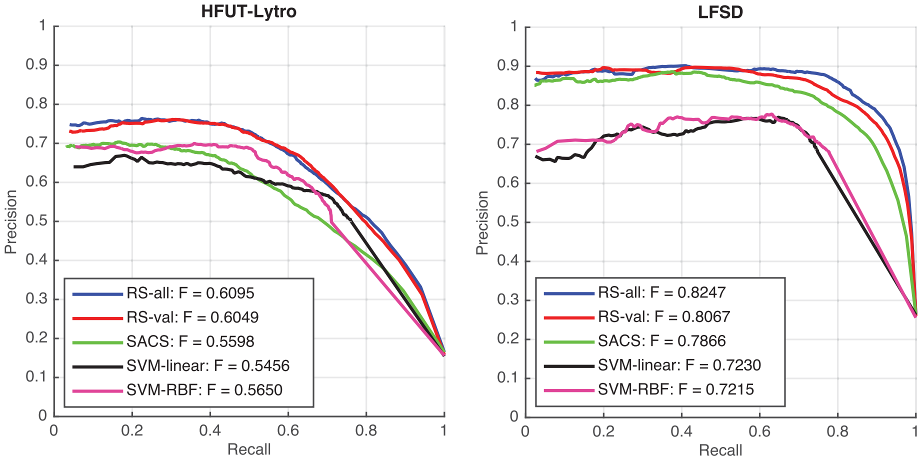

Fig. 11. Performance comparison of different multi-cue integration methods on the HFUT-Lytro and LFSD datasets. Here, we illustrate the PR curves and F-measures.

5.3.3. Evaluation on the Structure Cue. We evaluate the contribution of our graph-based refinement framework at different stages. Results are shown in Figure 10. From Figure 10(A), it is clear that the coarse saliency maps without refinement seem to capture the saliency location and global structural information (like shapes of the salient objects roughly), but the details, especially object boundaries and subtle structures, are easily lost and even mistakenly highlight some of the background regions particularly when the background is cluttered. Meanwhile, our approach with the second-stage refinement (R2) will lead to a better performance than only one stage (R1), especially on the challenging HFUT-Lytro dataset, as shown in Figures 10(B) and (C). Overall, our saliency refining strategy plays a crucial role in capturing the semantic object properties by introducing more spatial structural information.

5.3.4. On the Integration of Multiple Cues. To assess the benefit of the random search (RS) for multi-cue integration in the proposed scheme, we compare different fusion methods in Figure 11. Here, we introduce two other cue integration models as baselines: the adaptively weighted fusion strategy (SACS) [12] and the SVM-based learning mechanism [22]. In the RS method, we optimize the cue weights by using all the data (RS-all)

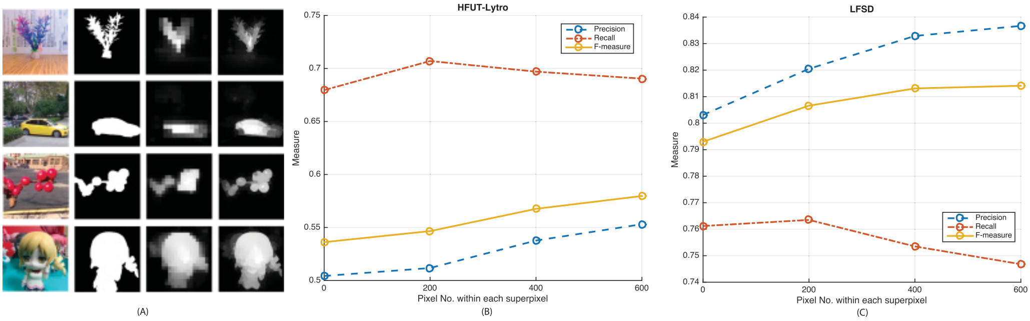

Fig. 12. Precision, recall, F-measure, and qualitative comparisons of saliency maps with and without SLIC. (A) Qualitative comparison of saliency maps. From left to right: all-focus images, ground truth, w/o SLIC, and w/ SLIC ($N = 400$); ((B) and (C)) Quantitative results on the HFUT-Lytro and LFSD datasets w.r.t. different numbers of pixels within each superpixel. 0 indicates regular grid sampling.

Table III. Computational Time of Our Approach and the State-of-the-art Methods for Processing One Image

Methods	Our approach	DILF [73]	WSC [29]	LFS [30]
Runtime (seconds)	4.2	0.9	8.5	8.1

or the validation set (RS-val). In the SVM method, we learn the model parameters using the validation set with linear (SVM-linear) or RBF (SVM-RBF) kernels to classify regions as salient vs. non-salient. We randomly sample a validation set (40% of the data in each dataset) for parameter learning. After this stage, the optimal parameters are computed to yield saliency maps on the remaining data. We repeat this procedure 3 times and report the average results.

It is fairly clear from Figure 11 that the random search method significantly outperforms other methods. In this case, the saliency model built on randomly selected parameters is able to better reflect the prior knowledge of any given scene and capture the saliency priors, which results in the good performance.

5.3.5. Evaluation on SLIC. To validate the effectiveness of SLIC, we simply replace SLIC with a regular grid sampling strategy and assign every light-field cue an equal weight without random search for fair comparison and measure the saliency maps generated with different numbers of pixels within each superpixel. Resulting comparisons are provided in Figure 12. It can be seen that the removal of SLIC leads to poor performance. This is not surprising, because regular grid segmentation ignores local structural constraints. The results also show that the setting of $N = 400$ leads to an overall better performance on the two datasets.

5.3.6. On the Computational Cost of the Approach. In terms of computation complexity, we compare the average runtime for each sample among different light-field saliency detection methods. We run the implementations by Matlab on an Intel i7 3.1GHz CPU PC with 16GB RAM. Table III shows the time cost of our approach compared with other state-of-the-art methods [29, 30, 73]. It can be seen that our approach consumes a smaller amount of computing time than the WSC [29] and LFS [30] and a bit more than DILF [73].

6. CONCLUSIONS

In this article, we propose a light-field saliency detection approach. In particular, we design a simple yet effective multi-cue scheme to encode the saliency priors in various visual channels, including color, depth, flow, and location. We investigate these visual cues in the context of a contrast-based saliency measurement where superpixels are used as direct inputs to the cue distinctiveness to detect the saliency presence of a superpixel. To integrate the saliency information from multiple perspectives, a random search method is employed to weight different saliency cues. Finally, we explore the structure cue to refine the object details and improve the accuracy of the saliency map. Our scheme does not involve any object detection or training process. To better analyze the effects of these visual cues and evaluate the proposed scheme, we collect light-field sequences of real-world scenes to construct a larger database with a set of light-field images and salient region annotations.

Extensive experimental comparisons have demonstrated that our approach significantly outperforms other advanced methods [29, 30, 73]. It yields a better performance by effectively exploring the complementation of multiple cues. Our multi-cue approach also offers novel insights for the understanding of light fields, which may potentially be useful in a wide variety of applications such as face detection [50], object recognition [70], and scene reconstruction [23].

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1597–1604.

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 11 (2012), 2274–2282.

[3] Edward H. Adelson and James R. Bergen. 1991. The plenoptic function and the elements of early vision. *Comput. Models Vis. Process.* 1, 2 (1991), 3–20.

[4] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. 2004. Interactive digital photomontage. *ACM Trans. Graph.* 23, 3 (2004), 294–302.

[5] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13, 1 (2012), 281–305.

[6] Ali Borji, Simone Frintrop, Dicky N. Sihite, and Laurent Itti. 2012. Adaptive object tracking by learning background context. In *Proceedings of the Computer Vision and Pattern Recognition Workshop on Egocentric Vision*. IEEE, 23–30.

[7] Ali Borji, Dicky N. Sihite, and Laurent Itti. 2012. Salient object detection: A benchmark. In *Proceedings of the European Conference on Computer Vision*. Springer, Berlin, Florence, Italy, 414–429.

[8] Ali Borji, Hamed R. Tavakoli, Dicky N. Sihite, and Laurent Itti. 2013. Analysis of scores, datasets, and models in visual saliency prediction. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 921–928.

[9] Thomas Brox, Andres Bruhn, Nils Papenberg, and Joachim Weickert. 2004. High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the European Conference on Computer Vision*, Vol. 4. Springer, Berlin, Prague, Czech Republic, 25–36.

[10] Neil Bruce and John Tsotsos. 2005. Saliency based on information maximization. In *Advances in Neural Information Processing Systems*. Curran Associates, Vancouver, Canada, 155–162.

[11] Neil Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K. Tsotsos. 2015. On computational modeling of visual saliency: Examining what's right, and what's left. *Vis. Res.* 116 (2015), 95–112.

[12] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng. 2014. Self-adaptively weighted co-saliency detection via rank constraint. *IEEE Trans. Image Process.* 23, 9 (2014), 4175–4185.

[13] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. 2008. Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems*. Curran Associates, Vancouver, Canada, 241–248.

[14] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. 2015. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 3 (March 2015), 569–582.

[15] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams. 2013. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1027–1034.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2011. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. Retrieved from http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[17] Tom Foulsham and Geoffrey Underwood. 2008. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *J. Vis.* 8, 2 (2008), 6–6.

[18] Alireza Ghasemi, Nelly Afonso, and Martin Vetterli. 2013. LCAV-31: A dataset for light-field object recognition. In *Proceedings of the International Society for Optics and Photonics Conference on Computational Imaging XII (SPIE 9020)*. SPIE, San Francisco, CA, 902014–902014.

[19] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. 2015. Accurate depth map estimation from a lenslet light field camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1547–1555.

[20] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. SALICON: Saliency in context. In *Proceedings of the Computer Vision and Pattern Recognition*. IEEE, 1072–1080.

[21] Tilke Judd, Frédo Durand, and Antonio Torralba. 2012. *A Benchmark of Computational Models of Saliency to Predict Human Fixations*. Technical Report. MIT.

[22] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2106–2113.

[23] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus H. Gross. 2013. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.* 32, 4 (2013), 73–1.

[24] Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. 2012. Depth matters: Influence of depth cues on visual saliency. In *Proceedings of the European Conference on Computer Vision*. Springer, Berlin, 101–115.

[25] Marc Levoy and Pat Hanrahan. 1996. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 31–42.

[26] Changyang Li, Yuchen Yuan, Weidong Cai, Yong Xia, and David Dagan Feng. 2015. Robust saliency detection via regularized random walks ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2710–2717.

[27] Guanbin Li and Yizhou Yu. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5455–5463.

[28] Jian Li, Martin D. Levine, Xiangjing An, Xin Xu, and Hangen He. 2013a. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 4 (2013), 996–1010.

[29] Nianyi Li, Bilin Sun, and Jingyi Yu. 2015. A weighted sparse coding framework for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5216–5223.

[30] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. 2014. Saliency detection on light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern RecognitionComputer Vision and Pattern Recognition*. IEEE, 2806–2813.

[31] Xi Li, Yao Li, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2013b. Contextual hypergraph modelling for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3328–3335.

[32] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. 2013c. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2976–2983.

[33] Chia-Kai Liang, Tai-Hsu Lin, Bing-Yi Wong, Chi Liu, and Homer H. Chen. 2008. Programmable aperture photography: Multiplexed light field acquisition. *ACM Trans. Graph.* 27, 3 (2008), 55.

[34] Ce Liu. 2009. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[35] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. 2015. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 362–370.

[36] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. 2011. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 2 (2011), 353–367.

[37] Yao Lu, Wei Zhang, Cheng Jin, and Xiangyang Xue. 2012. Learning attention map from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1067–1074.

[38] Andrew Lumsdaine and Todor Georgiev. 2009. The focused plenoptic camera. In *Proceedings of the International Conference on Computational Photography*. IEEE, 1–8.

[39] Ping Luo, Yonglong Tian, Xiaogang Wang, and Xiaoou Tang. 2014. Switchable deep network for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 899–906.

[40] Chih-Yao Ma and Hsueh-Ming Hang. 2015. Learning-based saliency model with depth information. *J. Vis.* 15, 6 (2015), 1–22.

[41] Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. 2013. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Trans. Graph.* 32, 4 (2013), 46.

[42] Hajime Mihara, Takuya Funatomi, Kenichiro Tanaka, and Hiroyuki Kubo. 2016. 4D light field segmentation with spatial and angular consistencies. In *Proceedings of the International Conference on Computational Photography*. IEEE, 1–8.

[43] Antoine Mousnier, Elif Vural, and Christine Guillemot. 2015. Partial light field tomographic reconstruction from a fixed-camera focal stack. *arXiv preprint arXiv:1503.01903* abs/1503.01903 (2015).

[44] Ren Ng. 2006. *Digital Light Field Photography*. Ph.D. Dissertation. Stanford University.

[45] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. 2005. *Light Field Photography with a Hand-held Plenoptic Camera*. Technical Report 2. Stanford University Computer Science.

[46] Nobuyuki Otsu. 1975. A threshold selection method from gray-level histograms. *Automatica* 11, 285–296 (1975), 23–27.

[47] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. 2014. RGBD salient object detection: A benchmark and algorithms. In *Proceedings of the European Conference on Computer Vision*. Springer, Berlin, 92–109.

[48] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 733–740.

[49] David Martin Powers. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* 2, 1 (2011), 37–63.

[50] Ramachandra Raghavendra, Kiran B. Raja, and Christoph Busch. 2014. Presentation attack detection for face recognition using light field camera. *IEEE Trans. Image Process.* 24, 3 (2014), 1060–1075.

[51] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. 2010. An eye fixation database for saliency detection in images. In *Proceedings of the European Conference on Computer Vision*, Vol. 6314. Springer, Berlin, 30–43.

[52] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. 2015. Exploiting global priors for RGB-D saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 25–32.

[53] Tongwei Ren, Yan Liu, Ran Ju, and Gangshan Wu. 2016. How important is location information in saliency detection of natural images. *Multimedia Tools Appl.* 75, 5 (2016), 2543–2564.

[54] Martin Rerabek and Touradj Ebrahimi. 2016. New light field image dataset. In *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE.

[55] Boris Schauerte and Rainer Stiefelhagen. 2013. How the distribution of salient objects in images influences salient object detection. In *Proceedings of the International Conference on Image Processing*. IEEE, 74–78.

[56] Atsushi Shimada, Hajime Nagahara, and Rin ichiro Taniguchi. 2013. Object detection based on spatio-temporal light field sensing. *IPSJ Trans. Comput. Vis. Appl.* 5, 0 (2013), 129–133.

[57] Michael W. Tao, Pratul P. Srinivasan, Jitendra Malik, Szymon Rusinkiewicz, and Ravi Ramamoorthi. 2015. Depth from shading, defocus, and correspondence using light-field angular coherence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1940–1948.

[58] Po-He Tseng, Ran Carmi, Ian G. M. Cameron, Douglas P. Munoz, and Laurent Itti. 2009. Quantifying center bias of observers in free viewing of dynamic natural scenes. *J. Vis.* 9, 7 (2009), 4.

[59] Vaibhav Vaish, Marc Levoy, Richard Szeliski, C. Lawrence Zitnick, and Sing Bing Kang. 2006. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. IEEE, 2331–2338.

[60] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. 2007. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.* 26, 3 (2007), 9.

[61] Eleonora Vig, Michael Dorr, and David Cox. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2798–2805.

[62] Junle Wang, Matthieu Perreira da Silva, Patrick Le Callet, and Vincent Ricordel. 2013. A computational model of stereoscopic 3d visual saliency. *IEEE Trans. Image Process.* 22, 6 (2013), 2151–2165.

[63] Ting-Chun Wang, Alexei Efros, and Ravi Ramamoorthi. 2015. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3487–3495.

[64] Sven Wanner, Stephan Meister, and Bastian Goldluecke. 2013. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modeling and Visualization*. The Eurographics Association, Lugano, Switzerland, 225–226.

[65] Gordon Wetzstein, Ivo Ihrke, Douglas Lanman, and Wolfgang Heidrich. 2011. Computational plenoptic imaging. *Comput. Graph. Forum* 30, 8 (2011), 2397–2426.

[66] Gordon Wetzstein, Douglas R. Lanman, Matthew Waggener Hirsch, and Ramesh Raskar. 2012. Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting. *ACM Trans. Graph.* 31, 4 (2012), 1–22.

[67] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. 2005. High performance imaging using large camera arrays. *ACM Trans. Graph.* 24, 3 (2005), 765–776.

[68] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. 2014. Predicting human gaze beyond pixels. *J. Vis.* 14, 1 (2014), 1–20.

[69] Linfeng Xu, Hongliang Li, Liaoyuan Zeng, and King Ngi Ngan. 2013. Saliency detection using joint spatial-color constraint and multiscale segmentation. *J. Vis. Commun. Image Represent.* 24, 4 (2013), 465–476.

[70] Yichao Xu, Kazuki Maeno, Hajime Nagahara, Atsushi Shimada, and Rin ichiro Taniguchi. 2015a. Light field distortion feature for transparent object classification. *Comput. Vis. Image Understand.* 139 (2015), 122–135.

[71] Yichao Xu, Hajime Nagahara, Atsushi Shimada, and Rin ichiro Taniguchi. 2015b. TransCut: Transparent object segmentation from a light-field image. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3442–3450.

[72] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3166–3173.

[73] Jun Zhang, Meng Wang, Jun Gao, Yi Wang, Xudong Zhang, and Xindong Wu. 2015. Saliency detection with a deeper investigation of light field. In *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI, 2212–2218.

[74] Qi Zhao and Christof Koch. 2012. Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. *J. Vis.* 12, 6 (2012), 1–15.

[75] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2015. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1265–1274.

[76] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3586–3593.

[77] Guokang Zhu, Qi Wang, and Yuan Yuan. 2014. Tag-saliency: Combining bottom-up and top-down information for saliency detection. *Comput. Vis. Image Understand.* 118 (2014), 40–49.

[78] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. 2014. Saliency optimization from robust background detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2814–2821.

Received September 2016; revised May 2017; accepted May 2017