Light Field Saliency Detection With Deep Convolutional Networks

Jun Zhang, Yamei Liu, Shengping Zhang, *Member, IEEE*, Ronald Poppe, and Meng Wang, *Member, IEEE*

***Abstract*— Light field imaging presents an attractive alternative to RGB imaging because of the recording of the direction of the incoming light. The detection of salient regions in a light field image benefits from the additional modeling of angular patterns. For RGB imaging, methods using CNNs have achieved excellent results on a range of tasks, including saliency detection. However, it is not trivial to use CNN-based methods for saliency detection on light field images because these methods are not specifically designed for processing light field inputs. In addition, current light field datasets are not sufficiently large to train CNNs. To overcome these issues, we present a new Lytro Illum dataset, which contains 640 light fields and their corresponding ground-truth saliency maps. Compared to current publicly available light field saliency datasets [1], [2], our new dataset is larger, of higher quality, contains more variation and more types of light field inputs. This makes our dataset suitable for training deeper networks and benchmarking. Furthermore, we propose a novel end-to-end CNN-based framework for light field saliency detection. Specifically, we propose three novel MAC (Model Angular Changes) blocks to process light field micro-lens images. We systematically study the impact of different architecture variants and compare light field saliency with regular 2D saliency. Our extensive comparisons indicate that our novel network significantly outperforms state-of-the-art methods on the proposed dataset and has desired generalization abilities on other existing datasets.**

***Index Terms*— Saliency detection, light field, micro-lens images, angular changes, deep neural network.**

I. Introduction

LIGHT field imaging [3] not only captures the color intensity of each pixel but also the directions of all incoming light rays. The directional information inherent in a light field implicitly defines the geometry of the observed scene [4].

Manuscript received April 11, 2019; revised October 8, 2019 and December 4, 2019; accepted January 22, 2020. Date of publication February 5, 2020; date of current version February 14, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61876057, Grant 61872112, Grant 61725203 and Grant 61732008, in part by the China Scholarship Council under Grant 201806695016, and in part by the National Key Research and Development Program of China under Grant 2018YFC0806802 and Grant 2018YFC0832105. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yannick Berthoumieu. *(Corresponding author: Jun Zhang.)*

Jun Zhang, Yamei Liu, and Meng Wang are with the Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230601, China, and also with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: zhangjun1126@gmail.com; liuyamei@mail.hfut.edu.cn; eric.mengwang@gmail.com).

Shengping Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China (e-mail: s.zhang@hit.edu.cn).

Ronald Poppe is with the Department of Information and Computing Sciences, Utrecht University, 3584 Utrecht, The Netherlands (e-mail: r.w.poppe@uu.nl).

Digital Object Identifier 10.1109/TIP.2020.2970529

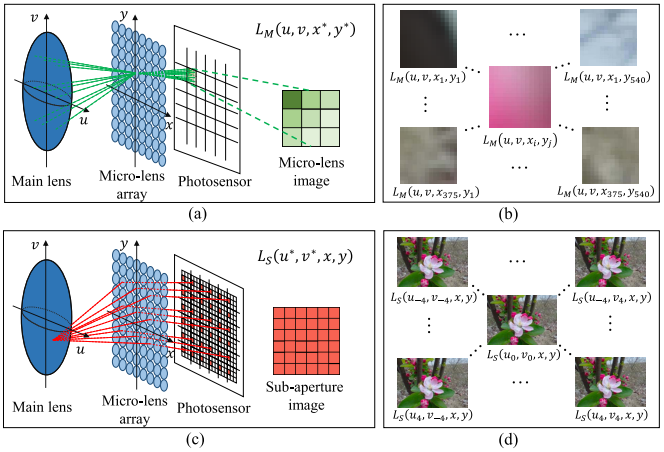

Fig. 1. Illustrations of light field representations. (a) Micro-lens image representation with the given location (x^*, y^*). (b) Micro-lens images at sampled spatial locations. (c) Sub-aperture image representation with the given viewpoint (u^*, v^*). (d) Sub-aperture images at sampled viewpoints, where (u_0, v_0) represents the central viewpoint.

In recent years, commercial and industrial light field cameras with a micro-lens array inserted between the main lens and the photosensor, such as Lytro [5] and Raytrix [6], have taken light field imaging into a new era. The obtained light field can be represented by 4D parameterization (u, v, x, y) [7], where uv denotes the viewpoint plane and xy denotes the image plane, as shown in Figures 1(a) and (c). The 4D light field can be further converted into multiple 2D light field images, such as multi-view sub-aperture images [7], micro-lens images [5], and epipolar plane images (EPIs) [8]. These light field images have been exploited to improve the performance of many applications, such as material recognition [9], face recognition [10], [11], depth estimation [12]–[16] and super-resolution [8], [17], [18].

This paper studies saliency detection on light field images. Previous work [1], [2], [19], [20] has focused on developing hand-crafted light field features at the superpixel level by utilizing heterogenous types of light field images (*e.g.,* color, depth, focusness, or flow). These methods strongly rely on low-level cues and are less capable of extracting high-level semantic concepts. This makes them unsuitable for handling highly cluttered backgrounds or predicting uniform regions inside salient objects.

In recent years, convolutional neural networks (CNNs) have been successfully applied to learn an implicit relation between pixels and salience in RGB images [21]–[27]. These CNN-based methods have been combined with object

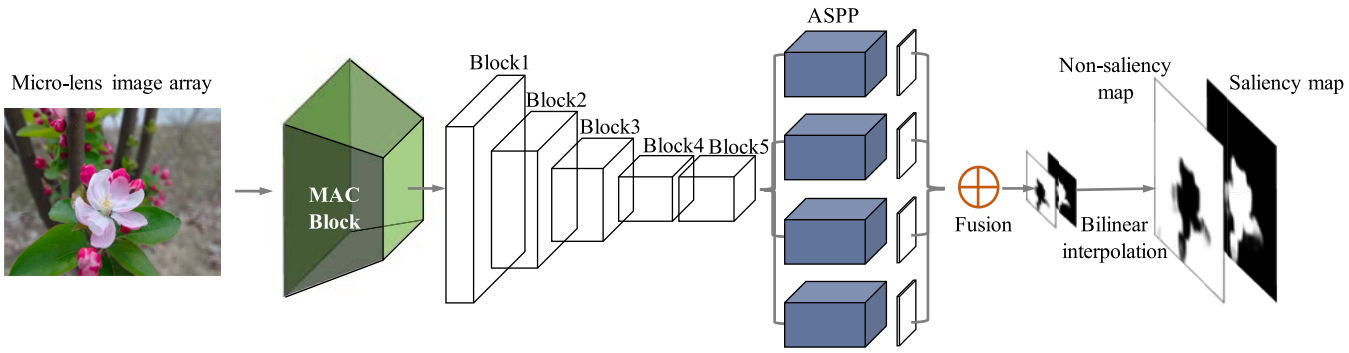

Fig. 2. Architecture of our network. The MAC building block converts the micro-lens image array of light fields into feature maps, which are processed by a modified DeepLab-v2 backbone model.

proposals [21], post-processing steps [22], contextual features [23], [24], attention models [25], [26], and recurrent structures [27]. Although these approaches achieve improved performance over saliency detection on benchmark datasets, they often adopt complex network architectures, which limits generalization and complicates training. Besides, the limited information in RGB images does not allow to fully exploit geometric constraints, which have been shown to be beneficial in saliency detection [28], [29].

Even with the emergence of CNNs for RGB images, there are still two key issues for saliency detection on light field images: (1) **Dataset.** The RGB datasets [30]–[32] are not sufficient to address significant variations in illumination, scale and background clutter. Previous publicly available light field saliency datasets LFSD [1] and HFUT-Lytro [2] include only 100 and 255 light fields, respectively, captured by the first-generation Lytro cameras. They are not large enough to train deep convolutional networks without severely overfitting. In addition, the unavailability of multi-views in the LFSD dataset and the color distortion of the sub-aperture images in the HFUT-Lytro dataset impede an evaluation of existing methods. (2) **Architecture.** The adoption of CNN-based architectures in light field saliency detection is not trivial because the existing CNNs for 2D images do not support the representation of 4D light field data. Thus, novel architectures must be developed for saliency detection in light field images.

There has been relatively little work on light field saliency detection using deep learning technologies. One recent and concurrent work [33] investigates different fusion structures to integrate the focal stack stream and the all-focus stream for saliency detection. To increase the diversity of the input data, they added adversarial examples to facilitate training and improving the robustness of the network. Besides, [33] builds a larger light field dataset for saliency detection, which is typically specialized on scenes with many focal slices located at different image depths. In this paper, we propose a novel method to predict the salience of light fields using CNNs. While both methods utilize deep learning technologies, ours is focused on the multi-view aspect based on micro-lens images, whereas their framework [33] is specifically designed for integrating light field data by using focal stacks and a recurrent network.

Specifically, to explore spatial and multi-view properties of light fields for saliency detection, we propose a novel deep convolutional network based on the modified DeepLab-v2 model [34] as well as several architecture variants (termed as *MAC* blocks) specifically designed for light field images. Our blocks aim to Model Angular Changes (MAC) from micro-lens images in an explicit way. One block type is similar to the angular filter explored in [9] for material recognition. The main difference is that [9] reorganizes the 4D information of the light field in different ways and divide the input into patches, which are further processed by the VGG-16 network for patch classification. We observe that this study does not pay much attention to the angular changes that may arise due to the network parameters, as it lacks an in-depth analysis of the relationship between its learned features and reflected information beneficial to material recognition. In contrast, inspired by the micro-lens array hardware configuration of light field cameras, the proposed MAC blocks are specially tailored to process micro-lens images in an explicit way. The network parameters are designed to sample different views and capture view dependencies by performing non-overlapping convolution on each micro-lens image. We experimentally show that the angular changes are consistent with the viewpoint variations of micro-lens images, and the effective angular changes of each pixel may increase depth selectivity and the ability for accurate saliency detection. Figure 2 provides an overview of the proposed network.

To train such a deep convolutional network, we introduce a comprehensive, realistic and challenging benchmark dataset for light field saliency detection. Using a Lytro Illum camera, we collect 640 light fields with significant variations in terms of size, amount of texture, background clutter and illumination. For each light field, we provide a high-quality micro-lens image array that contains multiple viewpoints for each spatial location. The micro-lens image array is firstly used as input for light field saliency detection. Then, we annotate per-pixel ground truth for each central viewing image.

Our contributions are summarized as follows:

- We construct a new light field dataset for saliency detection, which comprises of 640 high-quality light fields and the corresponding per-pixel ground-truth saliency maps. This dataset enables efficient deep network training

for saliency detection, and addresses new challenges in saliency detection such as inconsistent illumination and small salient objects in the cluttered or similar background.
- We propose an end-to-end deep convolutional network for predicting saliency on light field micro-lens images. To the best of our knowledge, no work has been reported on employing deep learning techniques for saliency detection to learn angular features from one single light field image.
- We provide an analysis of the proposed architecture variants specifically designed for light-field inputs. We also quantitatively and qualitatively compare our best-performing architecture with the 2D model using the central viewing image and other 2D RGB-based methods. We show that our network outperforms state-of-the-art methods on the proposed dataset and generalizes well to other datasets.

The remainder of this paper is structured as follows. The next section summarizes related work on light field datasets, saliency detection from light field images, and saliency detection using deep learning technologies. We introduce our novel Lytro Illum saliency dataset in Section III. We introduce our novel MAC blocks in Section IV and evaluate them in Section V. We conclude in Section VI.

II. Related Work

A. Light Field Datasets for Saliency Detection

There are only two existing datasets designed for light field saliency detection, both recorded with Lytro's first-generation cameras, which is capable of refocusing images after being taken. The Light Field Saliency Database (LFSD) [1] contains 100 light fields with 360×360 spatial resolution. A rough focal stack and an all-focus image are provided for each light field. The images in this dataset usually have one salient foreground object and a background with good color contrast. The limited complexity of the dataset is not sufficient to address the variety of challenges for saliency detection when using a light field camera, such as illumination variations and small objects on the similar or cluttered background. Later, Zhang *et al.* [2] proposed the HFUT-Lytro dataset, which consists of 255 light fields with complex backgrounds and multiple salient objects. Each light field has a 7×7 angular resolution and 328×328 pixels of spatial resolution. Focal stacks, sub-aperture images, all-focus images, and coarse depth maps are provided in this dataset. However, the color channels in their sub-aperture images are distorted owing to the under-sampling during decoding [35]. In this work, we use a second generation Lytro Illum camera to build a larger, higher-quality and more challenging saliency dataset by capturing more variations in illuminance, scale, and position. These two types of cameras differ in the number of microlenses and the number of pixels in the sensor beneath a microlens. Compared to the first-generation Lytro camera with a 11 megaray light-field sensor, the Lytro Illum comes equipped with a 40-megaray sensor. Therefore, the light field data obtained from a Lytro Illum camera have larger spatial resolution and angular resolution than those from the the first-generation Lytro camera. In addition, Lytro Illum's refocusing is finer and more granular, which allows to extend the refocusable range by capturing 3 or 5 consecutive images at different depths. In our work, we also generate the micro-lens image array from every decoded light field, which is not provided in previous datasets.

B. Saliency Detection on Light Field Images

Previous methods for light field saliency detection rely on superpixel-level hand-crafted features [1], [2], [19], [20], [36]. Pioneering work by Li *et al.* [1], [36] shows the feasibility of detecting salient regions using all-focus images and focal stacks from light fields. Zhang *et al.* [19] explored the light field depth cue in saliency detection, and further computed light field flow fields over focal slices and multi-view sub-aperture images to capture depth contrast [2]. In [20], a dictionary learning-based method is presented to combine various light field features using a sparse coding framework. Notably, these approaches share the assumption that dissimilarities between image regions imply salient cues. In addition, some of them [2], [19], [20] also utilize refinement strategies to enforce neighboring constraints for saliency optimization. In contrast to the above methods, we propose a deep convolutional network by learning efficient angular kernels without additional refinement on the upsampled image.

C. Deep Learning for Saliency Prediction

In the early days, saliency detection focused on using eye fixation locations as saliency maps, where the ground-truths are usually obtained by summing eye fixation binary maps of individual subjects and smoothed with a Gaussian of width dependent on the eye tracking set up. With the development of deep learning technologies, many deep networks are pre-trained on the ImageNet dataset [37], which makes the networks capable of identifying objects. Therefore, salient object detection has been attracting an increasing amount of research effort [21]–[27], thanks to the powerful representation learning methods. In these works, images are annotated salient objects with pixel-wise binary masks. In our work, we aim at detecting distinct salient objects/regions which attract human most.

Since the task is closely related to pixel-wise image classification, most works have built upon successful architectures for image recognition on the ImageNet dataset, often initializing their networks with the VGG network [38]. For example, several methods directly use CNNs to learn effective contextual features and combine them to infer saliency [23], [24]. Other methods extract features at multiple scales and generate saliency maps in a fully convolutional way [39], [40]. Recently, attention models [25], [26] have been introduced to saliency detection to mimic the visual attention mechanism by focusing on informative regions in visual scenes. Another direction for improving the quality of the saliency maps is the use of a recurrent structure [27], which mainly serves as a refinement stage to correct previous errors. Although deep CNNs have achieved great success in saliency detection, none of them addresses challenges in the 4D light field. Directly applying the existing network architectures to light field images would not be appropriate because a standard network

is not particularly good at capturing viewpoint changes in light fields.

D. Deep Learning Technologies on Light Field Data

Recently, in terms of different light field image types, learning-based techniques have been explored for light field image processing. Yoon *et al.* [17] proposed a deep learning framework for spatial and angular super-resolution, in which two adjacent sub-aperture images are employed to generate the in-between view. Wang *et al.* [41] built a bidirectional recurrent CNN to super-resolve horizontally and vertically adjacent sub-aperture image stacks separately and then combined them using a multi-scale fusion scheme to obtain complete view images. Very recently, Zhang *et al.* [42] designed a residual network structure to process one central view image and four stacks of sub-aperture images from four angular directions. Residual information from different directions is then combined to yield the high-resolution central view image. Kalantari *et al.* [43] proposed the first deep learning framework for view synthesis. They applied two sequential CNNs on only four corner sub-aperture images to model depth and color estimation simultaneously by minimizing the error between synthesized views and ground truth images. Wu *et al.* [18] introduced the "blur-restoration-deblur" framework for light field reconstruction on 2D EPIs (epipolar plane images). In order to directly synthesize novel views of dense 4D light fields from sparse views, Wang *et al.* [44] assembled 2D strided convolutions operated on stacked EPIs and two detail-restoration 3D CNNs connected with angular conversion to build a pseudo 4D CNN. Heber *et al.* [45] applied a CNN in a sliding window fashion for shape from EPIs, which allows to estimate the depth map of a predefined sub-aperture image. In the successive work [46], they designed a U-shaped network for disparity estimation operating on a EPI volume with two spatial dimensions and one angular dimension. Wang *et al.* [33] proposed the first and concurrent work on light field saliency using deep learning, in which they investigated different fusion structures to integrate the focal stack stream and the all-focus stream. To increase the diversity of the input data, they also used adversarial examples to facilitate training and improving the robustness of the network. Their work focuses on combining the light field data. In contrast, our approach is designed for learning angular features based on the micro-lens array hardware configuration, and we compare three different architectures using micro-lens images as inputs. The most similar to our work is [9], which proposes several CNN architectures based on different light field image types. One of the architectures is developed for the images similar to raw micro-lens images. However, the network is mainly designed to verify the advantages of multi-view information of light field compared with 2D RGB images in material recognition. In contrast, our work specifically focuses on the learned angular features from micro-lens images and their relationship with salient/non-salient cues.

III. The Lytro Illum Saliency Dataset

To train and evaluate our network for saliency detection, we introduce a comprehensive novel light field dataset.

A. Light Field Representation

There are various ways to represent the light field [7], [47], [48]. We adopt the two-plane parameterization [7] to define the light field as a 4D function $L(u, v, x, y)$, where $u \times v$ indicates the angular resolution and $x \times y$ indicates the spatial resolution. As illustrated in Figures 1(a) and (b), a set of all incoming rays from the uv plane intersected with a given micro-lens location (x^*, y^*) produces a micro-lens image with multiple viewpoints $L_M(u, v, x^*, y^*)$. The micro-lens images from different locations can be arranged into a micro-lens image array. As shown in Figures 1(c) and (d), all micro-lens regions on the xy plane receive the incoming rays from a given angular position (u^*, v^*), which produces a sub-aperture image with all locations $L_S(u^*, v^*, x, y)$. The central viewing image is formed by the rays passed through the main lens optical center $(u = u_0, v = v_0)$. Since the sub-aperture images contain optical distortions caused by the light rays passed through the lens [49], [50], in this paper, we build our network based on the micro-lens images, which have been shown advantages over the sub-aperture images for scene reconstruction [51].

B. Dataset Construction

Figure 3 illustrates the procedure of our light field dataset construction. First, a set of 4D light fields are obtained using a Lytro Illum camera (Figure 3(a)). Second, we use Lytro Power Tools (LPT) [52] to decode light fields from raw 4D data to 2D sub-aperture images so that each light field has a spatial resolution of 540×375 and an angular resolution of 14×14. To reach a compromise on the training time and the detection accuracy, we sample 9×9 viewpoints from each light field to generate new sub-aperture images, as shown in Figure 3(b). Third, we generate a micro-lens image by sampling the same spatial location from each sub-aperture image (see Figure 3(e)), which further produces a micro-lens image array of size 4860×3375, shown in Figure 3(c). The red region indicates one pixel with 9×9 observation viewpoints in Figure 3(c), comparing to one pixel only with the central view in Figure 3(d). We initially collect 800 light fields and manually annotate the per-pixel ground-truth label for each central viewing image. To reduce label inconsistency, each image is annotated by five independent annotators. We only regard a pixel as salient if it is verified by at least three annotators. We only keep those images with sufficient agreement. In the end, our new dataset contains 640 light fields with 81 views.

Figure 4 shows eight examples of central viewing images and their corresponding ground-truth saliency maps. There are significant variations in illumination, spatial distribution, scale and background. Besides, there are multiple regions for some saliency annotations.

IV. Light Field Saliency Network

We propose an end-to-end deep convolutional network framework for light field saliency detection as shown in Figure 2. Based on the micro-lens image array, the MAC (Modal Angular Changes) blocks are designed to transfer the light field inputs to feature maps in different ways. Then,

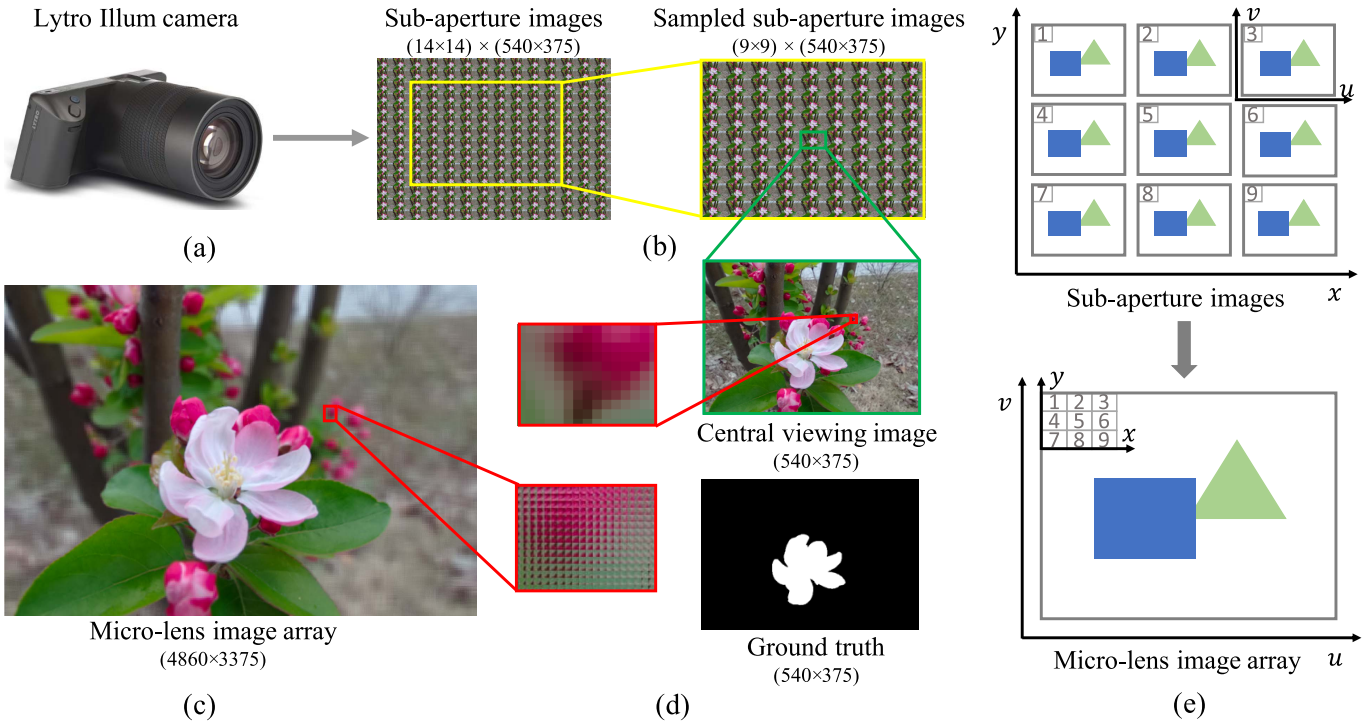

Fig. 3. Flowchart of the dataset construction. (a) Lytro Illum camera. (b) Sub-aperture images. (c) Micro-lens image array. (d) Ground-truth map for the central viewing image. (e) The generation of a micro-lens image array from sub-aperture images. The digits indicate viewpoints.

Fig. 4. Example central viewing images (top) and their corresponding ground-truth saliency maps (bottom) from our novel Lytro Illum dataset.

the feature maps are fed to a modified DeepLab-v2 [34] to predict saliency maps. We first discuss the backbone model and then detail different MAC block variants.

A. Backbone Model

We formulate light field saliency detection as a binary pixel labeling problem. Saliency detection and semantic segmentation are closely related because both are pixel-wise labeling tasks and require low-level cues as well as high-level semantic information. Inspired by previous literature on semantic segmentation [34], [53], [54], we design our backbone model based on DeepLab [54], which is a variant of FCNs [53] modified from the VGG-16 network [38]. There are several variants of DeepLab [34], [55], [56]. In this work, we use DeepLab-v2 [34], which introduces atrous spatial pyramid pooling (ASPP) to capture multi-scale information and long-range spatial dependencies among image units.

The modified network is composed of five convolutional (*conv*) blocks, each of which is divided into convolutions followed by a ReLu. A max-pooling layer is connected after the top *conv* layer of each *conv* block. The ASPP is applied on top of block5, which consists of four branches with atrous rates ($r = \{6, 12, 18, 24\}$). Each branch contains one 3×3 convolution and one 1×1 convolution. The resulting features from all branches are then passed through another 1×1 convolution and summed to generate the final score. The network further employs bilinear interpolation to upsample the fused score map to the original resolution of the central viewing image, which produces the saliency prediction at the pixel level. In addition, we add dropout to all the *conv* layers of the five blocks to avoid overfitting and set the 1×1 *conv* layer with 2 channels after ASPP to produce saliency and non-saliency score maps. The detailed architecture is illustrated in Figure 5.

B. MAC Blocks

Our network is essentially a modified DeepLab-v2 network augmented with a light field input process. As shown in Figure 2, a MAC block is a basic computational unit operating on a micro-lens image array input $\mathbf{M} \in \mathbb{R}^{W \times H \times C}$

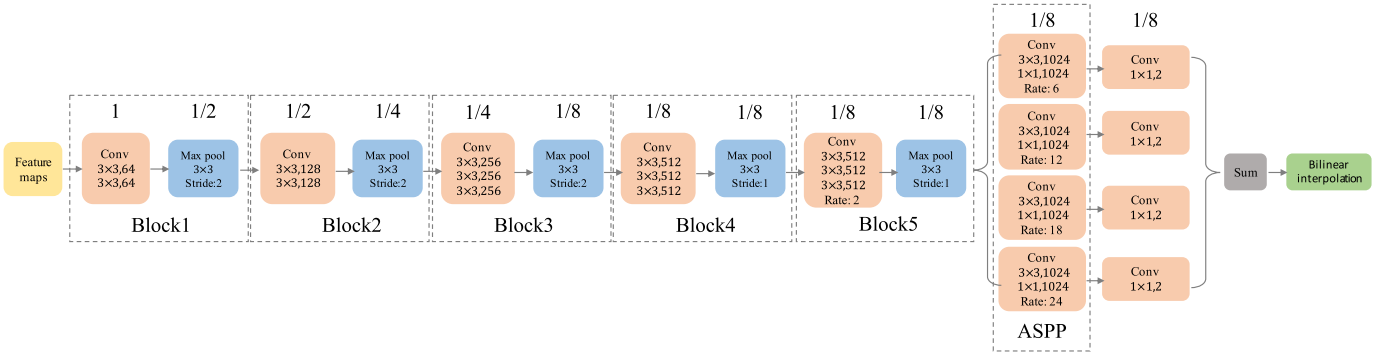

Fig. 5. Network structure of the backbone model based on DeepLab-v2 [34]. The reduction in resolution is shown at the top of each box.

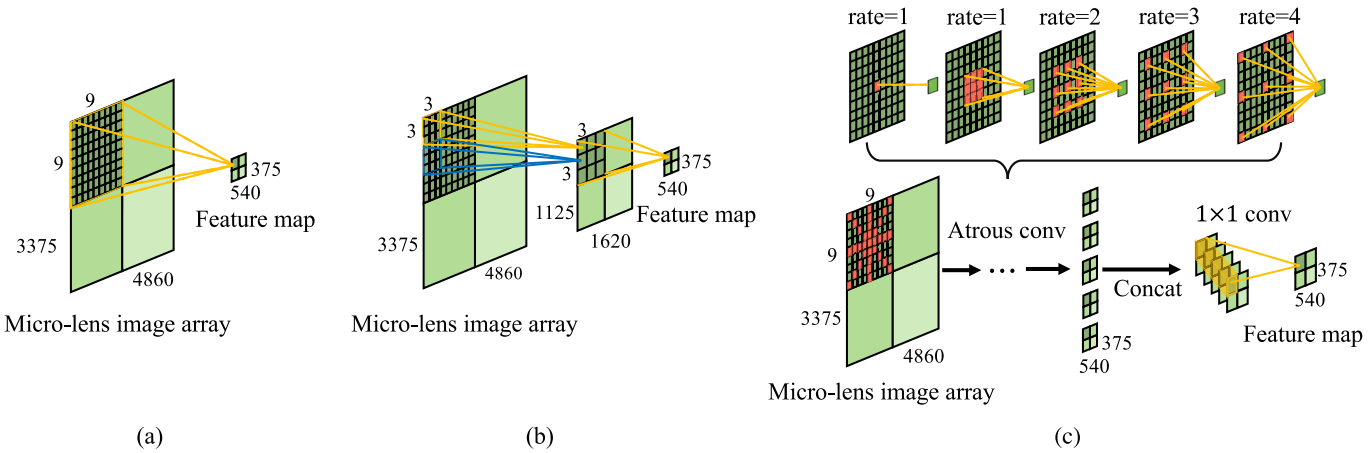

Fig. 6. Architectures of the proposed MAC blocks. (a) MAC block-9 × 9. (b) MAC block-3 × 3. (c) MAC block-star shaped. The selected viewpoints are highlighted in red.

and producing an output feature map $\mathbf{F} \in \mathbb{R}^{W' \times H' \times C'}$. Here, $W = N_x \times N_u$ and $H = N_y \times N_v$, in which (N_x, N_y) is the spatial size and (N_u, N_v) is the view size, respectively. The motivation of the MAC block is to model angular changes at one pixel location in an explicit manner. An essential part of learning angular features is the design of convolutional kernels applied on the micro-lens images. However, it is unclear what defines "good" angular filters and how many angular directions should be chosen for better performance. In this paper, we propose three different MAC block variant architectures to process light field micro-lens image arrays before block1 of the backbone model, in which convolutional methods with kernel sizes, stride size and sampled viewpoints are all designed to capture angular changes in light fields, as shown in Figure 6.

For the design simplicity of the MAC block, some default settings are fixed to guarantee that the predicted map and the ground truth map have the same spatial resolution in a fully convolutional network architecture. First, the spatial dimension of output of the MAC block is ensured to be the same with that of the 2D sub-aperture image, *i.e.* $W' = N_x$ and $H' = N_y$. Second, the number C' of convolutional kernels in the MAC block is the same as that of convolution kernels in block1 of DeepLab-v2. In our case where the data are captured by a Lytro Illum camera, the MAC block converts the light field input data into a $540 \times 375 \times 64$ feature map. The parameters of MAC block variants, including the kernel size $k \times k \times C$ and the convolutional stride s, should meet the above two conditions. We now discuss the detailed architectures of the three proposed MAC blocks.

1) MAC Block-9 × 9: As described in Section III-B, each micro-lens image has 9×9 viewpoints and can be considered as one of the pixel locations. The spatial resolution is 540×375 thus the size of the whole micro-lens image array is 4860×3375. In this architecture, we design angular convolutional kernels across all viewpoint directions, as shown in Figure 6(a). The kernel size shares the same angular resolution of one micro-lens image, and the number of kernels and the stride size are set to extract angular features for each micro-lens image. Specifically, we propose 64 angular kernels, each of which is a 9×9 filter. The stride of convolution operations is 9, which leads to $540 \times 375 \times 64$ feature maps. Each point on the feature map can be considered as being captured by the 81 lenslets. These kernels differ from common convolutional kernels applied on 2D images in that they only detect the angular changes in the micro-lens image array. This architecture directly learns the angular information from light field images, and thus is expected to distinguish salient foregrounds and backgrounds with similar colors or textures.

2) MAC Block-3 × 3: Motivated by the effectiveness of the smaller kernels in VGG-16 [38] and Inception v2 [55], we replace the 9×9 convolution in MAC block-9 × 9 with two

layers of 3×3 convolution (stride= 3) shown in Figure 6(b), which increases the number of parameters while enhancing the network nonlinearity.

3) MAC Block-Star Shaped: We design atrous angular convolutional kernels to capture long-range angular features. The atrous rates are set to sample representative viewpoint directions. It has been shown that using selected angular directions is beneficial in the context of depth estimation [14], [57]. Here, we test the application in saliency detection. Different from MAC block-9×9, we select star-shaped viewpoints (*i.e.* four directions $\theta = \{0°, 45°, 90°, 135°\}$) from each micro-lens image. To implement viewpoint sampling and angular filtering, we use atrous convolution with five atrous rates, as shown in Figure 6(c). The resulting feature maps are concatenated and combined using 1×1 convolutions for later processing.

Adaptation: Note that although the proposed framework is specially tailored to process micro-lens based light field data for saliency detection, in theory, the proposed network could be adapted to different types of light field data as well if the camera acquires the same pixel with sufficient different views. However, directly using the proposed network to process other types of light field data potentially has the following problems. (i) The number of views is usually limited and the resulting images suffer from angular aliasing due to the poor angular sampling of other cameras, such as multi-camera arrays [58] and light field gantries [59]. (ii) For sparse and wide-baseline light fields captured by multi-camera arrays, convolution operating over the full resolution of light fields may be prohibitively memory intensive and computationally expensive. (iii) Compared to micro-lens based light field cameras offered by simple design, flexibility and little marginal costs, other light field systems are generally impractical for the outdoor data collection due to the complex, heavy, and cost of the capturing system. It is difficult to construct a realistic and challenging saliency dataset of a certain scale. Therefore, the proposed network in this paper is currently only applicable to micro-lens based light field data with a set of dense views.

V. Experimental Results

Our experimental evaluation is split up into three main parts. The first section evaluates the three variations of the MAC blocks to identify the network design that works best for light-field saliency detection. The second section discusses the angular resolution, the overfitting issue and the advantages of the selected variant of the MAC block compared to 2D saliency detectors. Finally, the third section shows the performance we can attain based on the best performing model. We show state-of-the-art results on the new Lytro Illum, HFUT-Lytro [2] and LFSD datasets [1], based on a pre-trained network on the proposed dataset.

A. Settings

1) Implementation and Training: The computational environment has an Intel i7-6700K CPU@4.00GHz, 15GB RAM, and an NVIDIA GTX1080Ti GPU. We trained our network using the Caffe library [60] with the maximum iteration step of 160K. We initialize the backbone model with

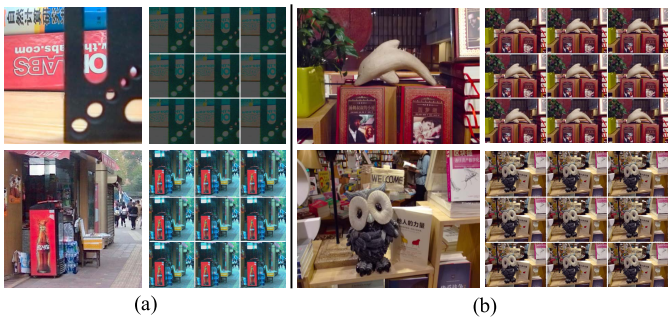

Fig. 7. Light field image examples. (a) The HFUT-Lytro dataset. (b) The proposed Lytro Illum dataset. Left: the all-focus images and the central viewing images are shown for the two datasets, respectively. Right: nine sub-aperture images are randomly sampled for each light field.

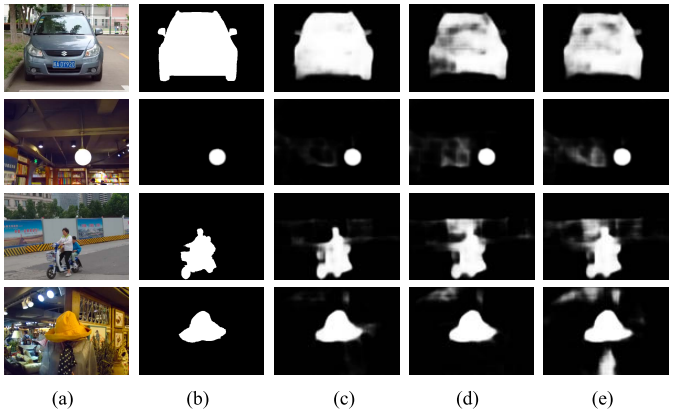

Fig. 8. Visual comparison of different MAC block variants. (a) Central viewing images. (b) Ground truth maps. (c) MAC block-9×9. (d) MAC block-3×3. (e) MAC block-star shaped.

DeepLab-v2 [34] pre-trained on the PASCAL VOC 2012 segmentation benchmark [61]. The newly added *conv* layers in the MAC block, the first layer of block1, and the score layer are initialized using the Xavier algorithm [62]. The whole network is trained end-to-end using the stochastic gradient descent (SGD) algorithm. To leverage the training time and the image size, we use a single image batch size. Momentum and weight decay are set to 0.9 and 0.0005, respectively. The base learning rate is initialized as 0.01 for the newly added *conv* layers in the MAC block and the first layer of block1, and 0.001 with the poly decay policy for the remaining layers. A dropout layer with probabilities $p = [0.1, 0.1, 0.2, 0.2, 0.3, 0.5]$ is applied after *conv* layers for block1–block5 and ASPP, respectively.

We use the softmax loss function defined as

$$L = -\frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} log \frac{e^{z_{i,j}^{y_{i,j}}}}{e^{z_{i,j}^{0}} + e^{z_{i,j}^{1}}} \tag{1}$$

where W and H indicate the width and height of an image, z_{ij}^{0} and z_{ij}^{1} are the last two activation values of the pixel (i, j) and y_{ij} is the ground-truth label of the pixel (i, j). Note that y_{ij} is 1 only when pixel (i, j) is salient. Our code and dataset are available at https://github.com/pencilzhang/MAC-light-field-saliency-net.git.

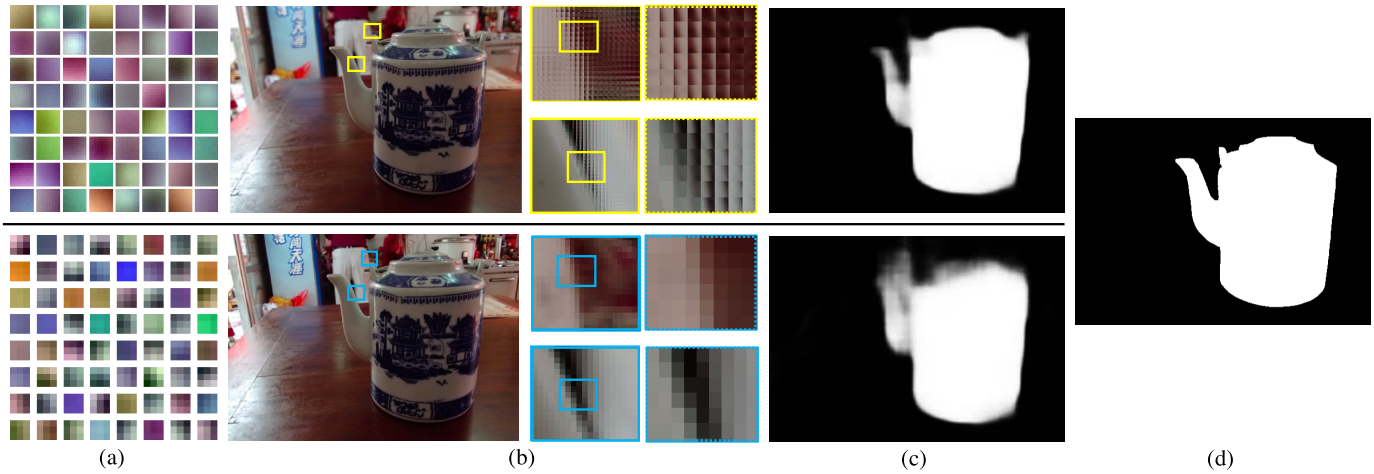

Fig. 9. Visual comparison of our 4D model (top) and 2D model using the central view (bottom). (a) Visualization of the first *conv* layers. (b) Light field input with highlighted regions. (c) Saliency predictions. (d) Ground truth maps.

2) Datasets: Three datasets are used for benchmarking: the proposed Lytro Illum dataset, the HFUT-Lytro dataset [2], and the LFSD dataset [1]. Our network is trained and evaluated on the proposed Lytro Illum dataset using a five-fold cross-validation. The trained model is further tested on the other two datasets to evaluate the generalization ability of our network. Note that the unavailable viewpoints in the LFSD dataset and the color distortion of sub-aperture images in the HFUT-Lytro dataset (see examples in Figure 7 for visual comparison) are unsuitable for evaluation of our method. To apply the trained model on the two datasets, we pad the angular resolutions to 9×9 using the all-focus image.

3) Data Augmentation: In order to obtain more training data to achieve good performance without overfitting, we augment the training data aggressively on-the-fly. To facilitate this augmentation, we use geometric transformations (*i.e.* rotation, flipping and cropping), changes in brightness, contrast, and chroma as well as additive Gaussian noise. Specifically, we rotate the micro-lens image array 90, 180, and 270 degrees, and perform horizontal and vertical flipping. To change the relative position of the saliency region in the image, we randomly crop two subimages of 3519×2907 size from the micro-lens image array. Then for one subimage and the image arrays with 0, 90, and 180 degrees of rotation, we adjust the brightness by multiplying all pixels by 1.5 and 0.6, respectively, and both chroma and contrast by the multiplication factor 1.7. Finally, we add the zero-mean Gaussian noise with variance of 0.01 to all images. In total, we expand the micro-lens image array by 48 ($(4 \times 4 + 8) \times 2$) such that the whole training dataset is increased from 512 to 24, 576.

4) Evaluation Metrics: We adopt five metrics to evaluate our network. The first one is precision-recall (PR) curve. Specifically, saliency maps are first binarized under varying thresholds and then compared to the ground truths. The second metric is F_β–measure, which considers both precision and recall

$$F_\beta = \frac{(1+\beta^2) Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (2)$$

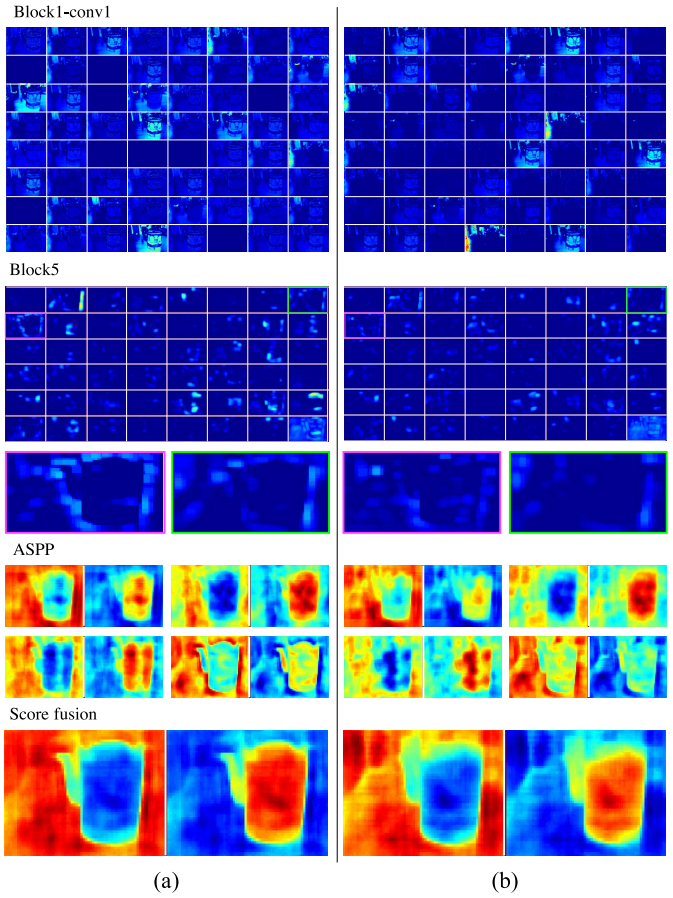

Fig. 10. Feature maps obtained from (a) 4D model and (b) 2D model using the central view from different layers. From top to bottom: the first *conv* features of block1, block5 output features, ASPP features with four atrous rates, and the score fusion maps via sum-pooling. For ASPP and sum fusion, the non-salience and salience scores are shown in the left and right subfigures, respectively.

where β^2 is set to 0.3 as suggested in [30]. Following previous work [22], [30], [31], we determine the adaptive threshold as twice the average value of the predicted saliency map to generate the binary map and report the corresponding

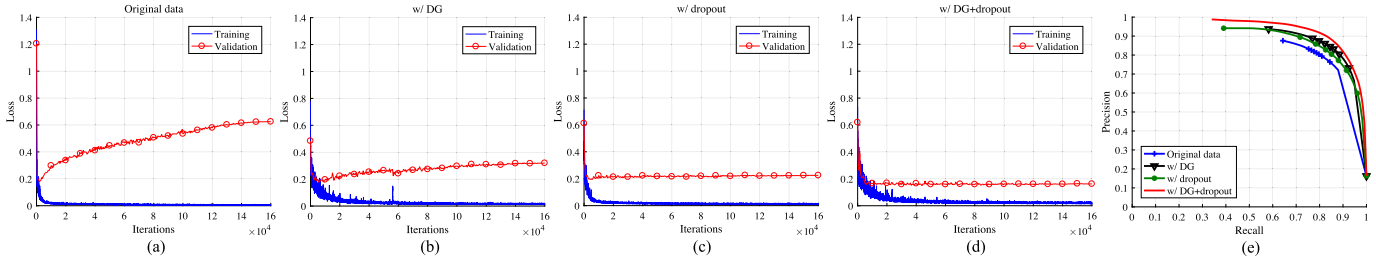

Fig. 11. Training and validation loss for our model on the proposed Lytro Illum dataset. (a) Original training data. (b) Training with DG. (c) Training with dropout. (d) Training with DG and dropout. (e) PR curves for different strategies.

TABLE I
QUANTITATIVE RESULTS ON THE PROPOSED LYTRO ILLUM DATASET

Method	F-measure	WF-measure	MAE	AP
MAC block-star shaped	0.8045	0.7426	0.0555	0.9120
MAC block-3 × 3	0.8066	0.7471	0.0562	0.9118
MAC block-9 × 9	**0.8116**	**0.7540**	**0.0551**	**0.9124**

mean F-measure value. The third metric is Average Precision (AP), which is computed by averaging the precision values at evenly spaced recall levels. The fourth metric is Mean Absolute Error (MAE), which directly computes the average absolute per-pixel difference between the predicted map and the corresponding ground truth map. Additionally, to amend several limitations of the above four metrics, such as interpolation flaw for AP, dependency flaw for PR curve and F_β–measure, and equal-importance flaw for all metrics, as suggested in [63], we use weighted F_β^w (WF)–measure based on weighted precision and recall as the fifth metric

$$F_\beta^w = \frac{(1+\beta^2)Precision^w \cdot Recall^w}{\beta^2 \cdot Precision^w + Recall^w} \tag{3}$$

where w is a weighting function based on the Euclidean distance to calculate the pixel importance from the ground truth.

B. Evaluation of MAC Blocks

We present a detailed performance comparison among different MAC block variant architectures on the proposed Lytro Illum dataset. As described in Section IV-B, these variants only differ in the convolution operations applied on their light field inputs. The quantitative results of the comparison are shown in Table I, from which we can see that the MAC block-9 × 9 architecture achieves the best performance for all metrics on the proposed dataset. We hypothesize that treating every micro-lens image as a whole and applying the angular kernels that have the same size with the angular resolution of the light field can help to exploit the multi-view information in the micro-lens image array. The detection performances of two other variants are lower, probably because the increased number of parameters make the network more difficult to train.

Figure 8 presents qualitative results of all variants. As illustrated in the figure, these variants can separate the most salient regions from similar or cluttered backgrounds. Compared to other variants, MAC block-9 × 9 outputs cleaner and more consistent predictions for the regions with specular reflections (row 1), small salient objects (row 2), and similar foreground and background (rows 2 and 3). Moreover, we can see that MAC block-9 × 9 better predicts salient regions without being highly affected by the light source (row 4). These results demonstrate that the proposed network variants are likely to extract potential depth cues by learning angular changes, which are helpful to saliency detection. The kernels with the same size of the angular resolution show better capability in depth discrimination.

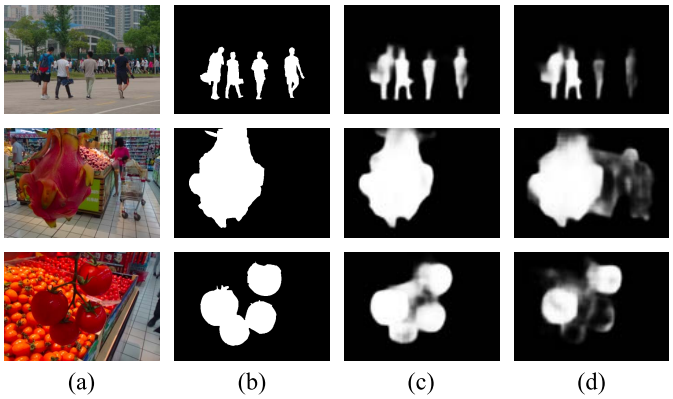

Fig. 12. Qualitative comparison of 4D model and 2D-central view. (a) Central viewing images. (b) Ground-truth maps. (c) Ours. (d) 2D-central view.

C. Model Analysis

Here, we perform all following experiments using MAC block-9 × 9, since this setup performed best in previous evaluation.

1) Effectiveness of the MAC Block: To further delve into the difference between regular image saliency and light field saliency, we present some important properties of light field features that can better facilitate saliency detection. We compare our 4D light field saliency (*i.e.* MAC block-9 × 9) to 2D model using the central viewing image as input (2D-central view). The quantitative results are shown in Table II. We observe that light field saliency detection with multi-views turns out to perform better than the 2D detector with only the central view.

To provide complementary insight of why light field saliency works, we visualize the weights of the first *conv* layers of our network and 2D-central view in Figure 9(a)

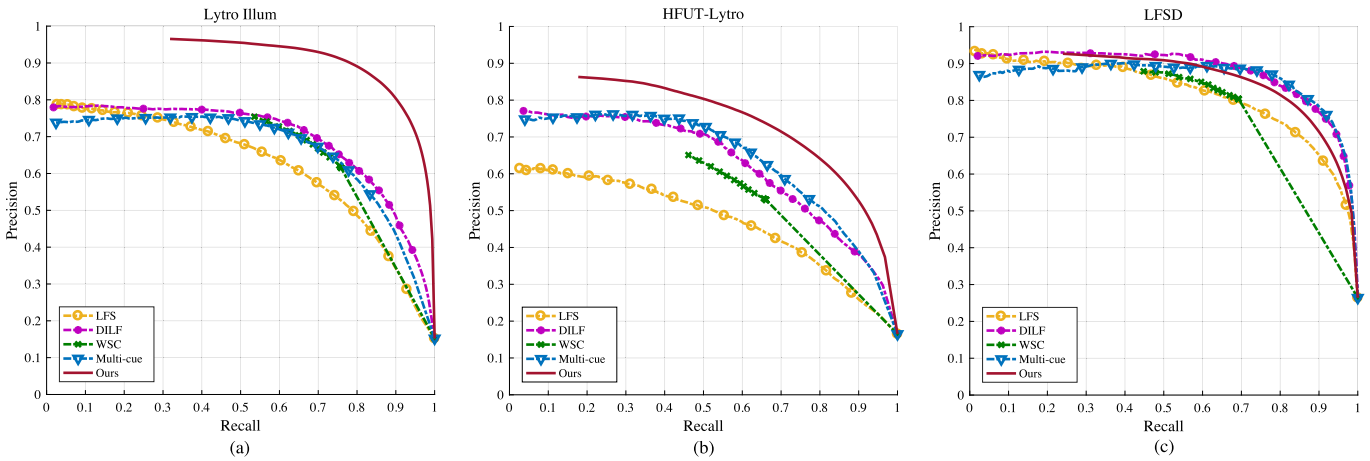

Fig. 13. Comparison on three datasets in terms of PR curve. (a) The proposed Lytro Illum dataset. (b) The HFUT-Lytro dataset. (c) The LFSD dataset.

TABLE II
QUANTITATIVE COMPARISON BETWEEN OUR 4D MODEL AND 2D-CENTRAL VIEW ON THE PROPOSED LYTRO ILLUM DATASET

Method	F-measure	WF-measure	MAE	AP
Ours	**0.8116**	**0.7540**	**0.0551**	**0.9124**
2D-central view	0.8056	0.7446	0.0597	0.9016

to compare angular and spatial patterns. We can see that the learned weights from our MAC block have noticeable changes in angular space, which suggests that the viewpoint cue of light field data is well captured. The angular changes are also consistent with the viewpoint variations of micro-lens images, as shown in Figure 9(b). The results are attributed to the newly designed *conv* method in which the kernel size is the same as the angular resolution of the micro-lens image, and the stride length guarantees angular features are extracted for each micro-lens image. Therefore, our 4D saliency detector produces more accurate saliency maps than the 2D detector shown in Figure 9(c).

In addition, we show the feature maps obtained from the two models in Figure 10. It can be seen that different layers encode different types of features. Higher layers capture semantic concepts of the salient region, whereas lower layers encode more discriminative features for identifying the salient region. The proposed 4D saliency detector can well discriminate the white spout from the white pants, as shown in Figure 10(a). However, as illustrated in the block1-conv1 and block5 of Figure 10(b), most feature maps from the 2D detector have small values that are not discriminative enough to separate the salient tea cup from the pants. Thus the 2D detector produces features cluttered with background noise in the following ASPP and score fusion. More comparisons of saliency maps between the two models can be seen in Figure 12.

2) Effect of the Angular Resolutions: To show the effect of the angular resolutions in the network, we compare the performance of our architecture with varying number of viewpoints in Table IV. Note that we change the kernel size to stay the same with the angular resolution. From the table, we can see that the network using 9×9 viewpoints shows the best performance overall. Increasing the angular resolution to 11×11 cannot improve the performance, which can be explained by the fact that the viewing angles at the boundary are very oblique [71] and the narrow baseline of the light field camera leads to high viewing redundancy with higher angular resolutions [7], [72].

3) Overfitting Issues: Overfitting is a common problem related to training a CNN with limited data. In this section, we analyse the proposed network by introducing different strategies to handle overfitting: data augmentation (DG) and dropout. The results obtained for our best performing model are shown in Figure 11. Clearly, the network is overfitting with original training data as shown in Figure 11(a). As expected, both DG and dropout are crucial to minimize overfitting as shown in Figures 11(b)–(d). Figure 11(e) presents the corresponding PR curves. It can be seen that by increasing the amount and diversity of the data and the amount of dropout between different layers during training, the performance of the network increases as well.

D. Comparison With 2D Models

To understand the additional information contained in the micro-lens light field images, we compare our best performing approach (*i.e.* MAC block-9×9) to 8 existing methods on the test set of our proposed dataset. Our comparison includes 4 traditional approaches MST [64], SMD [65], MDC [66], WFD [67]; and 4 CNN-based ones: PiCANet [24], Amulet [68], LFR [69], HyperFusion [70]. To facilitate fair comparison and effective model training, we use the recommended parameter settings provided by the authors to initialize these models. All CNN-based methods are based on DNNs pre-trained on the ImageNet [37] classification task. We retrain these CNN models on the proposed dataset in a five-fold cross-validation way and apply the same data augmentation method used in our work. The quantitative results are shown in Table III.

We can see that in general, our model outperforms other methods in terms of F-measure, MAE, and AP metrics. Amulet [68] obtains the second best performance on the

Fig. 14. Visual comparison of our best MAC block variant (Ours) and state-of-the-art methods on three datasets. (a) Central viewing/all-focus images. (b) Ground truth maps. (c) Ours. (d) LFS [36]. (e) DILF [19]. (f) WSC [20]. (g) Multi-cue [2]. The first five samples are taken from the proposed Lytro Illum dataset, the middle three samples are taken from the HFUT-Lytro dataset, and the last two samples are taken from the LFSD dataset.

proposed dataset. CNN-based methods consistently perform better than traditional methods. Additionally, we found that our whole network learning consumes less GPU memory compared to most compared 2D deep learning saliency methods.

E. Comparison to State-of-the-Art Light Field Methods

We compare our best performing model MAC block-9 $\times$ 9 to four state-of-the-art methods tailored to light field saliency detection: Multi-cue [2], DILF [19], WSC [20], and LFS [1]. We train our network on the novel dataset, and evaluate on the

TABLE III

QUANTITATIVE COMPARISON OF OUR APPROACH AND OTHER 2D MODELS ON THE PROPOSED DATASET. BOLD: BEST, UNDERLINED: SECOND BEST

	Model	F-measure	WF-measure	MAE	AP
Traditional	MST [64]	0.6695	0.5834	0.1243	0.6967
	SMD [65]	0.7246	0.5371	0.1234	0.7789
	MDC [66]	0.7407	0.5891	0.1094	0.7552
	WFD [67]	0.7260	0.6408	0.1024	0.7604
CNN-based	PiCANet [24]	0.7908	0.6745	0.0782	0.8362
	Amulet [68]	0.8059	**0.7686**	0.0552	0.8485
	LFR [69]	0.7756	0.7242	0.0702	0.8463
	HyperFusion [70]	0.7549	0.6945	0.0752	0.8359
	Ours	**0.8116**	0.7540	**0.0551**	**0.9124**

TABLE IV

EFFECTS OF THE ANGULAR RESOLUTION ON THE PROPOSED DATASET

Angular resolution	F-measure	WF-measure	MAE	AP
7×7	0.8018	0.7406	0.0567	**0.9135**
9×9	**0.8116**	**0.7540**	**0.0551**	0.9124
11×11	0.8006	0.7392	0.0567	0.9109

TABLE V

QUANTITATIVE RESULTS ON THE PROPOSED LYTRO ILLUM DATASET. BOLD: BEST, UNDERLINED: SECOND BEST

Method	F-measure	WF-measure	MAE	AP
LFS [1]	0.6107	0.3596	0.1697	0.6193
WSC [20]	0.6451	0.5945	0.1093	0.5958
DILF [19]	0.6395	0.4844	0.1389	0.6921
Multi-cue [2]	0.6648	0.5420	0.1197	0.6593
Ours	**0.8116**	**0.7540**	**0.0551**	**0.9124**

TABLE VI

QUANTITATIVE RESULTS ON THE HFUT-LYTRO DATASET. BOLD: BEST, UNDERLINED: SECOND BEST

Method	F-measure	WF-measure	MAE	AP
LFS [1]	0.4868	0.3023	0.2215	0.4718
WSC [20]	0.5552	0.5080	0.1454	0.4743
DILF [19]	0.5543	0.4468	0.1579	0.6221
Multi-cue [2]	0.6135	0.5146	0.1388	0.6354
Ours	**0.6721**	**0.6087**	**0.1029**	**0.7390**

others without fine-tuning. The results of other methods are obtained using the authors' implementations. Tables V–VII and Figure 13 show quantitative results on three datasets. Overall, our approach outperforms other methods on three datasets without any post-processing for refinement, which demonstrates the advantage of the proposed deep convolutional network for light field saliency detection. In particular, we observe that the proposed approach shows significant performance gains when compared to previous methods on the proposed dataset for all metrics. The performance is lower on the HFUT-Lytro and LFSD datasets, which is due to the limited viewpoint information in these datasets. Therefore, a large number of filters learnt on the proposed dataset are underused. This demonstrates that different light field datasets do affect the accuracy of methods. Multi-cue [2] and DILF [19] methods show better performance than our approach in terms of F-measure and AP on the LFSD dataset. The reason is that these methods use external depth features and post-processing refinement to improve the performance.

TABLE VII

QUANTITATIVE RESULTS ON THE LFSD DATASET. BOLD: BEST, UNDERLINED: SECOND BEST

Method	F-measure	WF-measure	MAE	AP
LFS [1]	0.7525	0.5319	0.2072	0.8161
WSC [20]	0.7729	0.7371	0.1453	0.6832
DILF [19]	0.8173	0.6695	0.1363	**0.8787**
Multi-cue [2]	**0.8249**	0.7155	0.1503	0.8625
Ours	0.8105	**0.7378**	**0.1164**	0.8561

Some qualitative results are shown in Figure 14. We can see that our approach can handle various challenging scenarios, including multiple salient objects (rows 1 and 2), cluttered backgrounds (rows 3 and 5), small salient objects (rows 4 and 7), inconsistent illumination (rows 1 and 6), and salient objects in similar backgrounds (rows 8, 9 and 10). It is also worth noting that without any post-processing, our approach can highlight salient objects more uniformly than other methods.

VI. CONCLUSION

This paper introduces a deep convolutional network for saliency detection on light fields by exploiting multi-view information in micro-lens images. Specifically, we propose MAC block variants to process the micro-lens image array. To facilitate training such a deep network, we introduce a challenging saliency dataset with light field images captured from a Lytro Illum camera. In total, 640 high quality light fields are produced, making the dataset more suitable for deep network training. Extensive experiments demonstrate that comparing to 2D saliency based on the central view alone, 4D light field saliency can exploit additional angular information contributing to an increase in the performance of saliency detection. The proposed network is superior to saliency detection methods designed for 2D RGB images on the proposed dataset, and outperforms the state-of-the-art light field saliency detection methods on the proposed dataset and generalizes well to the existing datasets. In particular, our approach is capable of detecting salient regions in challenging cases, such as with similar foregrounds and backgrounds, inconsistent illumination, multiple salient objects, and cluttered backgrounds. Our work suggests promising future directions of exploiting spatial and angular patterns in light fields and deep learning technologies to advance the state-of-the-art in pixel-wise prediction tasks.

REFERENCES

[1] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1605–1616, Aug. 2017.

[2] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui, "Saliency detection on light field: A multi-cue approach," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 3, pp. 1–22, 2017.

[3] E. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[4] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 41–48.

[5] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Stanford Univ. Comput. Sci., Stanford, CA, USA, Tech. Rep. CSTR2, 2005.

[6] Raytrix GmbH. Accessed: 2010. [Online]. Available: https://raytrix.de/

[7] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH*, 1996, pp. 31–42.

[8] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[9] T.-C. Wang, J.-Y. Zhu, and E. Hiroaki, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 121–138.

[10] R. Raghavendra, K. B. Raja, and C. Busch, "Presentation attack detection for face recognition using light field camera," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1060–1075, Mar. 2015.

[11] A. Sepas-Moghaddam, M. A. Haque, P. L. Correia, K. Nasrollahi, T. B. Moeslund, and F. Pereira, "A double-deep spatio-angular learning framework for light field based face recognition," 2018, *arXiv:1805.10078v2*. [Online]. Available: https://arxiv.org/abs/1805.10078v2

[12] M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, and R. Ramamoorthi, "Shape estimation from shading, defocus, and correspondence using light-field angular coherence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 546–560, Mar. 2017.

[13] Williem, I. K. Park, and K. M. Lee, "Robust light field depth estimation using occlusion-noise aware data costs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2484–2497, Oct. 2018. doi: 10.1109/TPAMI.2017.2746858.

[14] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4748–4757.

[15] H. Schilling, M. Diebold, C. Rother, and B. Jähne, "Trust your model: Light field depth estimation with inline occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4530–4538.

[16] H.-G. Jeon *et al.*, "Depth from a light field image with learning-based matching costs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 297–310, Feb. 2019, doi: 10.1109/TPAMI.2018.2794979.

[17] Y. Yoon, H.-G. Jeon, and D. Yoo, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 24–32.

[18] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on EPI and extended applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1681–1694, Jul. 2019.

[19] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu, "Saliency detection with a deeper investigation of light field," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2212–2218.

[20] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5216–5223.

[21] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3183–3192.

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

[23] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1265–1274.

[24] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[25] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3668–3677.

[26] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 714–722.

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

[28] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.

[29] V. Sitzmann *et al.*, "Saliency in VR: How do people explore virtual environments?" *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 4, pp. 1633–1642, Apr. 2018.

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

[31] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[32] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[33] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, "Deep learning for light field saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8838–8848.

[34] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[35] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.

[36] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp 2806–2813.

[37] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

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

[40] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. V. Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5781–5790.

[41] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4274–4286, Sep. 2018.

[42] S. Zhang, Y. Lin, and H. Sheng, "Residual networks for light field image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11046–11055.

[43] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, p. 193, 2016.

[44] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, "End-to-end view synthesis for light field imaging with pseudo 4DCNN," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 333–348.

[45] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3746–3754.

[46] S. Heber, W. Yu, and T. Pock, "Neural EPI-volume networks for shape from light field," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2252–2260.

[47] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proc. Comput. Graph. Interact. Techn.*, 1995, pp. 39–46.

[48] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. Cambridge, MA, USA: MIT Press, 1991.

[49] H. Tang and K. N. Kutulakos, "What does an aberrated photo tell us about the lens and the scene?" in *Proc. IEEE Int. Conf. Comput. Photogr.*, Apr. 2013, pp. 1–10.

[50] H.-G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1547–1555.

[51] S. Zhang, H. Sheng, D. Yang, J. Zhang, and Z. Xiong, "Micro-lens-based matching for scene recovery in lenslet cameras," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1060–1075, Mar. 2018.

[52] *Lytro Power Tools*. Accessed: 2018. [Online]. Available: https://github.com/kmader/lytro-power-tools

[53] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[54] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2016, *arXiv:1412.7062v4*. [Online]. Available: https://arxiv.org/abs/1412.7062v4

[55] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[56] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*. [Online]. Available: https://arxiv.org/abs/1802.02611

[57] M. Strecke, A. Alperovich, and B. Goldluecke, "Accurate depth and normal maps from occlusion-aware focal stack symmetry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2529–2537.

[58] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.

[59] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," *Vis., Model., Vis.*, vol. 13, pp. 225–226, Sep. 2013.

[60] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," 2014, *arXiv:1408.5093*. [Online]. Available: https://arxiv.org/abs/1408.5093

[61] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[62] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[63] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[64] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2334–2342.

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

[66] X. Huang and Y.-J. Zhang, "300-FPS salient object detection via minimum directional contrast," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4243–4254, Sep. 2017.

[67] X. Huang and Y. Zhang, "Water flow driven salient object detection at 180 fps," *Pattern Recognit.*, vol. 76, pp. 95–107, Apr. 2018.

[68] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 202–211.

[69] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection by lossless feature reflection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1149–1155.

[70] P. Zhang, W. Liu, Y. Lei, and H. Lu, "Hyperfusion-Net: Hyper-densely reflective feature fusion for salient object detection," *Pattern Recognit.*, vol. 93, pp. 521–533, Sep. 2019.

[71] M. Levoy, Z. Zhang, and I. Mcdowall, "Recording and controlling the 4D light field in a microscope using microlens arrays," *J. Microscopy*, vol. 235, no. 2, pp. 144–162, Aug. 2009.

[72] M. Le Pendu, X. Jiang, and C. Guillemot, "Light field inpainting propagation via low rank matrix completion," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1981–1993, Apr. 2018.

Jun Zhang received the B.S., M.S., and Ph.D. degrees from the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), China, in 2007, 2009, and 2013, respectively. From August 2010 to September 2012, and May 2015 to September 2015, she worked at Brown University as a visiting Research Fellow. She also acts as an Associate Professor with the School of Computer Science and Information Engineering, HFUT, from December 2015. From July 2013 to December 2017, she was a Postdoctoral Fellow in computer science at HFUT. From November 2018 to November 2019, she was a Visiting Researcher at Utrecht University. Her current interests include computer vision, vision perception, and computation photography.

Yamei Liu received the B.S. degree from the School of Computer Science and Information Engineering, Hefei University of Technology, China, in 2017, where she is currently pursuing the master's degree with the School of Computer Science and Information Engineering. Her research interests include computer vision, image processing and analysis, and machine learning.

Shengping Zhang (Member, IEEE) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He had been a Postdoctoral Research Associate at Brown University and with Hong Kong Baptist University, and a Visiting Student Researcher with the University of California at Berkeley. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai. He has authored or coauthored over 50 research publications in refereed journals and conferences. His research interests include deep learning and its applications in computer vision. He is also an Associate Editor of the *Signal Image and Video Processing*, and the *Journal of Electronic Imaging*.

Ronald Poppe received the Ph.D. degree in computer science from the University of Twente, The Netherlands, in 2009. He was a Visiting Researcher at the Delft University of Technology, Stanford University, and University of Lancaster. He is currently an Assistant Professor with the Information and Computing Sciences Department, Utrecht University. His research interests include modelling of visual attention and the analysis of human (interactive) behavior from videos and other sensors. In 2012 and 2013, he received the Most Cited Paper Award from Image and Vision Computing. In 2017, he received a TOP Grant from the Dutch Science Foundation.

Meng Wang (Member, IEEE) received the B.E. and Ph.D. degrees in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, China. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored more than 200 book chapters, journals, and conference papers in these areas. He is a recipient of the ACM SIGMM Rising Star Award 2014. He is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (IEEE TKDE), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (IEEE TCSVT), the IEEE TRANSACTIONS ON MULTIMEDIA (IEEE TMM), and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (IEEE TNNLS).