Acquiring Dynamic Light Fields through Coded Aperture Camera

Kohei Sakai[1][0000−0001−9059−1171], Keita Takahashi[1][0000−0001−9429−5273], Toshiaki Fujii[1][0000−0002−3440−5132], and Hajime Nagahara[2][0000−0003−1579−8767]

[1] Graduate School of Engineering, Nagoya University, Japan
{sakai.kohei, takahasi, fujii}@fujii.nuee.nagoya-u.ac.jp
[2] Institute for Datability Science, Osaka University, Japan,
nagahara@ids.osaka-u.ac.jp

Abstract. We investigate the problem of compressive acquisition of a dynamic light field. A promising solution for compressive light field acquisition is to use a coded aperture camera, with which an entire light field can be computationally reconstructed from several images captured through differently-coded aperture patterns. With this method, it was assumed that the scene should not move throughout the complete acquisition process, which restricted real applications. In this study, however, we assume that the target scene may change over time, and propose a method for acquiring a dynamic light field (a moving scene) using a coded aperture camera and a convolutional neural network (CNN). To successfully handle scene motions, we develop a new configuration of image observation, called V-shape observation, and train the CNN using a dynamic-light-field dataset with pseudo motions. Our method is validated through experiments using both a computer-generated scene and a real camera.

Keywords: Light Field, CNN, Coded Aperture Camera

1 Introduction

The concept of a light field, which is a 4-D signal representation that describes all light rays traveling in 3-D space [1,12,23], has been used in various applications, such as view synthesis [19,27,34,54], depth estimation [35,44,48,51], synthetic refocusing [18,32], super resolution [5,48], 3-D display [16,22,39,49], and object recognition [25,45]. A light field is usually represented as a set of dense multi-view images, where many (tens to hundreds) views are aligned in parallel with tiny viewpoint intervals.

Acquiring a light field is challenging due to the large amount of data, for which several approaches have been investigated. The most straightforward approach is to use a moving camera gantry [23] or multiple cameras [11, 38, 50] to capture a target scene from different viewpoints. This approach is costly in terms of the hardware or time required to acquire the entire light field. Another approach is to use lens-array based cameras that can capture both the spatial

and directional information of the light rays [2, 3, 31, 32]. These cameras can acquire an entire light field in a single image, but the spatial resolution of each viewpoint image is in a trade-off relationship with the number of viewpoints.[3] The final approach we mention is compressed acquisition using, e.g., a coded aperture/mask camera [4, 17, 24, 26, 30, 40, 43, 53]. We are interested in the final approach due to its potential advantage in efficiency and the ability to acquire a light field in the full spatial resolution of the image sensor.

With this final approach, the target light field is computationally reconstructed from several observed images of the same target scene with different encoding (aperture/mask) patterns. [4] The number of images required for reconstruction can be successfully reduced by optimizing the encoding process (e.g., finding optimal aperture/mask patterns) and corresponding reconstruction algorithm. In earlier studies, this problem was tackled in the context of compressed sensing [6, 7, 10], and reconstruction methods were developed on the basis of sparse representation on a learned dictionary [26, 40] and approximation using the most significant basis vectors [53]. In more recent studies, deep neural networks were successfully applied for better reconstruction from fewer observed images [13, 17, 29, 42]. For example, Inagaki et al. [17] reported that only a few observed images were sufficient for reconstructing a light field with 5×5 or 8×8 views. This successful result came with the learning-based optimization of the entire acquisition process modeled using a deep convolutional neural network (CNN). However, most of the methods mentioned here have been applied only to static light fields (stationary scenes).

In this study, we focused on the problem of compressive acquisition of dynamic light fields (moving scenes). Specifically, we extended Inagaki et al.'s method [17], which was designed exclusively for static light fields, to dynamic light fields. In short, we propose a method for acquiring a dynamic light field using a coded aperture camera and a CNN. To our knowledge, this is the first work that achieves compressive acquisition of a dynamic light field based on the concept of "deep optics", where the optical elements (aperture patterns) and reconstruction algorithm are jointly optimized through deep learning.

Given successful results [17] for static light fields, one might easily conceive of the following two naive strategies for dynamic light fields, both of which are unsuccessful, as shown from our experiments. The first strategy is to reconstruct a light field at each time from only a single observed image [26], which helps avoid the effect of scene motions. Thanks to the recent deep-learning-based optimization, the quality of light field reconstruction from a single image has improved [8, 29, 37]. However, it is essentially difficult for this strategy to achieve

[3] The combination of a lens-array based camera and ordinary camera has also been explored to increase the temporal resolution [46], but the trade-off between the spatial and directional resolutions still remains unsolved.

[4] A related topic is angular super resolution [19, 27, 47, 52, 54], where the target light field is synthesized from sparser (e.g., located at the four corners) views. This is regarded as a special case of compressive acquisition where the encoding process is limited to view subsampling.

geometrically-correct reconstruction, because in principle, the disparity information cannot be extracted from a single observed image alone. In particular, when an ordinary image (without aperture/mask coding) is used as the input [8, 37], the resulting light field is only "hallucinated" based on the implicit knowledge learned from the training dataset rather than the apparent geometric cues. Another naive strategy is to assume that the scene is stationary for a short time and directly apply a method designed for static light fields. In this case, a light field at a certain time can be reconstructed from several images observed over different times [40]. Using several images helps in obtaining 3-D information embedded as disparities related to different aperture patterns. This strategy was expected to work well for scenes with few motions. However, it fails in practice because scene motions are non-negligible even between two consecutive times.

To summarize, we need several (at least two) images to obtain sufficient geometric cues for reconstructing a 3-D structure of the target scene, which are embedded as disparity information among the images captured through different aperture patterns. At the same time, the observed images are also affected by the scene motions over time. In other words, the difference among the observed images is not only due to the disparities but also by scene motions, which greatly complicates the reconstruction problem. To address this issue in our method, we first introduced a new configuration of image observation, called V-shape observation, to help the CNN successfully separate the disparity information from scene motions. We then constructed a dynamic-light-field dataset from static light fields with pseudo motions, and used it for training the CNN to make the CNN more adaptable to dynamic scenes. Our method was quantitatively evaluated through simulation experiments using a computer-generated dynamic scene. We also applied our method for a coded aperture camera and succeeded in capturing a real dynamic scene with fine quality.

2 Proposed Method

We present a method for acquiring a dynamic light field (a moving scene) with a coded aperture camera and a CNN. Our method can be regarded as an extension of the method by Inagaki et al. [17] that was designed exclusively for static light fields. However, to our knowledge, our work is the first to achieve compressive acquisition of a dynamic light field based on the concept of deep optics, where the optical elements (aperture patterns) and reconstruction algorithm are jointly optimized for dynamic light fields through deep learning.

In this section, we first introduce notations and the problem formulation in Section 2.1. Next, we explore several possible configurations for dynamic light-field reconstruction and discuss the proposed method in Section 2.2. We explain the architecture of the CNN in Section 2.3 then describe the datasets, training procedure, and the optimized aperture patterns in Section 2.4.

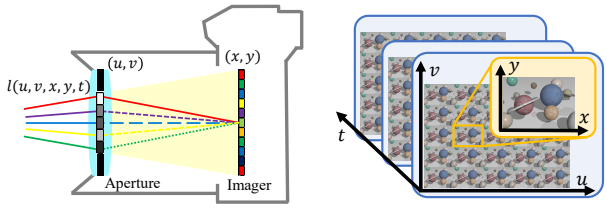

Fig. 1. Coded aperture camera (left) and example of dynamic light field (right)

2.1 Notations and Problem Formulation

A schematic diagram of a coded aperture camera is shown in Fig. 1 (left). All incoming light rays to the camera are parameterized with four variables (u, v, x, y), where (u, v) and (x, y) denote the intersections with the aperture and imaging planes, respectively. The light field is defined over 4-D space (u, v, x, y), with which the light intensity is described as $l(u, v, x, y)$. When we consider scene motions over time t, the light intensity is described as $l(u, v, x, y, t)$ on 5-D space.

We consider a coded aperture design with which the transmittance of the aperture can be controlled at any position and time. The transmittance at position (u, v) and t is defined as $a(u, v, t)$. The image-formation process through a coded aperture camera is described as

$$i_t(x, y) = \frac{1}{|\mathcal{E}_t|} \iiint_{\mathcal{E}_t \times \mathcal{U} \times \mathcal{V}} a(u, v, \tau) l(u, v, x, y, \tau) du dv d\tau, \tag{1}$$

where $i_t(x, y)$ is the observed image at t, $\mathcal{E}_t$ is the exposure time around t, and $\mathcal{U} \times \mathcal{V}$ denotes the effective aperture area.

We next transform Eq. 1 into a discretized representation on the time and aperture domains. We assume that the light field and aperture pattern are constant during each exposure time, i.e., $l(u, v, x, y, \tau) = l_t(u, v, x, y)$ and $a(u, v, \tau) = a_t(u, v)$ for $\tau \in \mathcal{E}_t$. In this case, t can be considered as an index instead of a real value. We also assume that the aperture plane is discretized into several square blocks indexed by a pair of integers (u, v). We can simplify Eq. 1 as

$$i_t(x, y) = \sum_{u,v} a_t(u, v) l_{u,v,t}(x, y). \tag{2}$$

We rewrite $l_t(u, v, x, y)$ as $l_{u,v,t}(x, y)$, which can be regarded as one of the rectified sub-aperture images observed from viewpoint (u, v) at t. Figure 1 (right) illustrates a case in which the aperture plane was discretized in 5×5 regions; thus, a light field at each t is represented as 5×5 multi-view images. The observed image given by Eq. 2 is a weighted sum of those multi-view images.

Given the model of Eq. 2, our goal is to reconstruct the original light field at each t, $l_{u,v,t}(x, y)$, from several observed images around t: $i_{t'}(x, y)$ for $t' \in \mathcal{T}_t$, where $\mathcal{T}_t$ denotes the local temporal window around t. The aperture patterns,

$a_t(u, v)$, should also be optimized simultaneously. This is a problem of compressed sensing with an extreme compression ratio, where a set of multi-view images (e.g., 5×5 views) is compressed into a single observed image at each t. However, in the reconstruction stage, we can use the information not only from the corresponding time ($i_t(x, y)$) but also from other adjacent times $t' \in \mathcal{T}_t$ (e.g., $i_{t-1}(x, y)$ and $i_{t+1}(x, y)$), which will help improve reconstruction quality.

The observation and reconstruction processes of a dynamic light field can be translated into a neural network model. The observation process at t, which is given by Eq. 2, can be written in a form of a mapping as $f\colon L_t \to I_t$ where L_t represents a tensor that contains all the pixels of $l_{u,v,t}(x, y)$ for all viewpoints (u, v) for a specific t, and I_t represents a tensor that contains all the pixels of $i_t(x, y)$ at t. The reconstruction process is written as $g\colon \{I_{t'} | t' \in \mathcal{T}_t\} \to \hat{L}_t$ where $\hat{L}_t$ corresponds to an estimate of L_t. The composite mapping $h = g \circ f$ can be regarded as an auto-encoder, where f and g correspond to the encoder and decoder, respectively, and a set of observed images, $i_{t'}(x, y)$ for $t' \in \mathcal{T}_t$, is regarded as a latent representation. The goal of optimization is formulated, e.g., with the squared error loss, as

$$\hat{f}, \hat{g} = \arg\min_{f,g} ||L_t - \hat{L}_t||^2. \tag{3}$$

As detailed later, we implemented the composite mapping as a deep CNN, using 2-D convolutional neural layers exclusively. The entire network can be trained end-to-end by using a training dataset. The learned parameters in $\hat{f}$ and $\hat{g}$ correspond to the aperture patterns $a_t(u, v)$ and reconstruction algorithm, respectively, both of which are jointly optimized. When applied to a real coded aperture camera, $\hat{f}$ is conducted by the physical imaging process on the camera and the aperture patterns of which are configured in accordance with the learned parameters in $\hat{f}$. The images acquired from the camera are fed to the network corresponding to $\hat{g}$, then, the target light field is reconstructed on the computer.

Our problem described above is similar but more challenging than that of Inagaki et al. [17]. In [17], the target light field is assumed to be static and the observation process is described as

$$i_t(x, y) = \sum_{u,v} a_t(u, v) l_{u,v}(x, y). \tag{4}$$

Time t still appears in $a_t(u, v)$ and $i_t(x, y)$ but disappears from $l_{u,v}(x, y)$. In this case, the same light field can be observed several times as several images $i_t(u, v)$ observed through different aperture patterns $a_t(u, v)$ over t. The target light field was reconstructed with reasonable fidelity because the difference in the observed images was caused solely by the difference in the aperture patterns. More intuitively, due to the difference in the masking patterns on the aperture plane, the observed images have disparities in accordance with the depth of each pixel (x, y), from which the reconstruction algorithm can deduce 3-D information of the target scene. In our case, however, the target light field $l_{u,v,t}(x, y)$ changes over t; thus, each light-field instance $l_{u,v,t}(x, y)$ can be observed only once. Similarly to Inagaki et al. [17], we change the aperture patterns over t, but the

differences in the observed images are due to not solely by the difference in the aperture patterns, which is known and even controllable, but also by the scene motions, which are unknown and should be estimated from the observed images.

2.2 Reconstruction of Dynamic Light Field

As mentioned earlier, we can use several observed images $I_{t'}$ ($t' \in \mathcal{T}_t$) to reconstruct a light field $\hat{L}_t$ at t. We now discuss how to do this more specifically considering several design factors. We then present our method, V-shape observation trained with our dynamic-light-field dataset.

The first factor is the number of aperture patterns used for observation. According to Inagaki et al. [17], only two images observed through two different aperture patterns are sufficient to reconstruct a static light field consisting of 5×5 or 8×8 views. Therefore, we determined to use at most two aperture patterns; using more patterns would be helpful to improve the reconstruction quality, but we did not do this to avoid increasing complexity. As shown at the bottom of Fig. 2, two aperture patterns, A and B, are alternately repeated over t. Therefore, at each t, we have only one observed image with one of the aperture patterns. An image observed through aperture pattern A at t is denoted as I_t^A. Note that the target light field changes over t.

The second and third factors are the number of observed images used for reconstruction and the type of training data: static or dynamic light fields. The possible reconstruction methods we considered along with the proposed method are summarized in the top-left table of Fig. 2 and discussed in detail below.

(i) **Single**: reconstruction from only a single observed image (top-right in Fig. 2). At each t, L_t is reconstructed from a single observed image I_t^A, using a pre-trained decoder $\hat{g}$, denoted as "CNN" in the figure. In this case, $\mathcal{T}_t = \{t\}$; thus, the model is free from scene motions; the training can be conducted with a static dataset, and the reconstruction is not affected by scene motions. However, it is difficult to expect good reconstruction quality because the disparity information cannot be obtained from a single image alone in principle.

(ii) **2-S**: reconstruction from two consecutive images using a model trained on a static dataset (bottom-left in Fig. 2). This is a naive application of Inagaki et al.'s method [17] for reconstructing a dynamic light field. We assume that the scene is static over t in which two images are captured and simply apply the model trained on a static dataset. We adopt $\mathcal{T}_t = \{t-1, t\}$ and try to reconstruct L_t from I_{t-1}^A and I_t^B. One might expect that this would work well with little motion because Inagaki et al.'s method [17] worked perfectly for static scenes. However, from our experiments, the scene motion cannot be negligible even between two consecutive images, which leads to poor reconstruction quality.

(iii) **2-D**: reconstruction from two consecutive images using a model trained on a dynamic dataset. This is the same as (ii) except for the training dataset; the model is trained on a dynamic dataset, which will make the model more adaptable to dynamic scenes. However, even in this case, the reconstruction quality is insufficient. One possible reason is that the scene motion and disparity information are inseparable on the two observed images. As mentioned earlier,

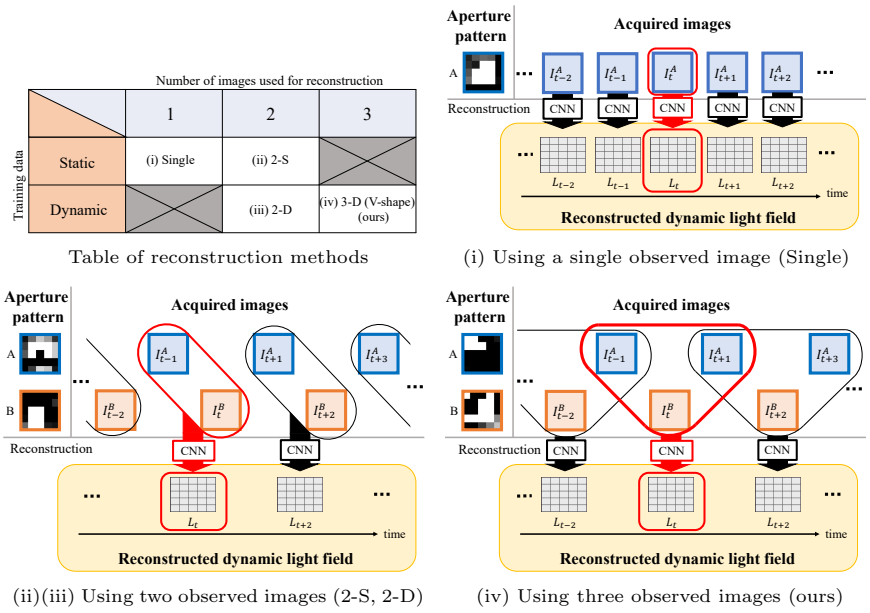

Table of reconstruction methods

(i) Using a single observed image (Single)

(ii)(iii) Using two observed images (2-S, 2-D)

(iv) Using three observed images (ours)

Fig. 2. Reconstruction of dynamic light field with several possible configurations

the difference between the two images is caused by the difference in the aperture patterns (which induces disparities) and scene motions.

(iv) **3-D (V-shape)**: reconstruction from three consecutive images using a model trained on a dynamic dataset (bottom right in Fig. 2). This is our proposed method. We adopt $\mathcal{T}_t = \{t-1, t, t+1\}$ and try to reconstruct L_t from the three observed images, I^A_{t-1}, I^B_t, and I^A_{t+1}. Images I^A_{t-1} and I^A_{t+1} are captured with the same aperture pattern, i.e., A, so that the difference between these images is exclusively attributed to scene motions. Image I^B_t contains both disparity and motion information with respect to the other two images. We expect that feeding these three images can help the CNN successfully separate disparity information from scene motions, which leads to better reconstruction quality of L_t. We call this "V-shape" observation because the locus tracing the three images constitutes a "V" shape.

2.3 Network Architecture

Figure 3 illustrates an example of the networks we constructed where the light field is composed of 25 (5×5) viewpoints and the temporal window is set to $\mathcal{T}_t = \{t-1, t, t+1\}$. The basic architecture is similar to that of Inagaki et al.'s [17], which was dedicated to static light fields, but our network can handle dynamic light fields thanks to the extensions described in Section 2.2. We now briefly summarize the architecture.

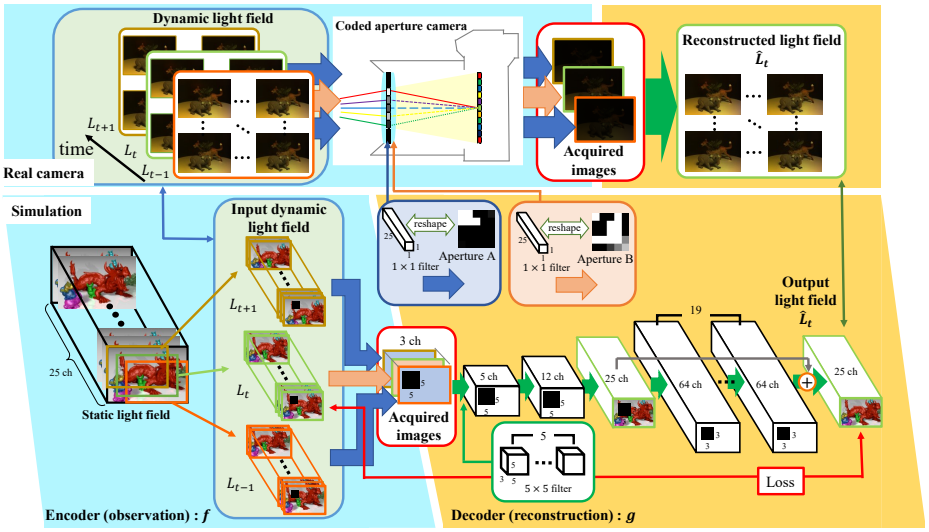

Fig. 3. Network structure for dynamic light field acquisition, where encoder and decoder parts are shown in blue and orange background, respectively.

The network is composed solely of 2-D convolutional layers; thus, it is agnostic of the spatial resolution. An instance of a light field at t (L_t) is treated as a 2-D image having multiple channels. More specifically, a light field with 25 viewpoints is translated into a 2-D image with 25 channels[5]. Therefore, the input and output of the network have 25 channels. For each convolution layer, we use one-pixel stride and appropriate zero padding to maintain the image size before and after the convolution. The number of channels can be changed at each convolution layer. Therefore, the image size (height and width) is kept constant throughout the network but only the number of channels is changed as the data proceed in the network.

The encoder part of the network f is designed with some constraints because it corresponds to the physical-image-capturing process of a coded aperture camera. Specifically, the transmittance of the aperture pattern ($a_t(u, v)$) should be limited within $[0, 1]$, and each pixel (x, y) on the imaging plane cannot be mixed with its neighbors. Similarly to Inagaki et al. [17], we implemented this process using a single 2-D convolutional layer that has a 1×1 convolution kernel and reduces the number of channels from 25 to 1 (the weights of this layer correspond to the transmittance of the aperture pattern). To keep the range limitation for the kernel weights, we clipped the weights within $[0, 1]$ each time when the network was updated with a mini-batch during the training stage. The bias terms were kept to 0. In the training stage, we added zero-mean Gaussian noise to I_t, which is important to make the learned model robust against camera noise.

[5] We assume that the light field has only one color channel for simplicity. For a color light field, RGB color channels are treated individually.

The decoder part of the network g can take an arbitrary form because the whole process is executed on the computer. We use the same decoder as that adopted by Inagaki et al. [17]. This decoder was designed to gradually increase the number of channels from three to 25 to obtain a tentative output then refine the tentative output through another deep residual CNN developed for image super-resolution [20]. The channel-increase step consists of several convolutional layers with 5×5 kernels and linear activation, and the refinement step consists of 19 convolution layers with 3×3 kernels and rectified-linear-unit activation.

2.4 Dataset and Training Procedure

We first explain how we prepared static and dynamic light field datasets used to train the networks. We then mentioned the details of the training procedure and the obtained aperture patterns. The number of views for a light field was set to 5×5 and 8×8.

We followed the procedure of Inagaki et al. [17] in preparing the static-light-field dataset. The training samples were collected from many light-field datasets [9, 14, 15, 28], which are summarized in Table 1. From each light field, we extracted image patches with 64×64 pixels at the same position from all 25 or 64 views and combined them to compose a light-field sample. The position of the image patches was changed to obtain many samples from each light field; we took patches every 32 pixels both in the horizontal and vertical directions but discarded those with almost uniform intensities. Three (RGB) color channels of each light field were used as three individual light fields. We augmented the data by changing the intensity levels of each sample uniformly. We multiplied 1.0, 0.9, 0.8, 0.7, 0.6 and 0.5 with the original samples. Finally, we collected 295,200 and 166,680 samples for 5×5 and 8×8 views, respectively, which were used to train the networks for Single and 2-S.

We next prepared a dynamic light field dataset, which was necessary to train the networks for 2-D and 3-D (V-shape). We need a dataset in which each sample consists of 5×5 or 8×8 views over three consecutive times. To the best of our knowledge, there are no public datasets suitable for our purpose. Therefore, we created such a dataset from the static-light-field dataset by giving it pseudo motions. As illustrated in Fig. 4, we clipped out three slightly different regions from a single image patch and regarded them as a temporal sequence. More specifically, we extracted three image patches with 60×60 pixels from each of the patches (64×64 pixels) of the static-light-field dataset. The extracted patches, gathered from all the views, constituted a sample of dynamic light field corresponding to a set of L_{t-1}, L_t and L_{t+1}. We assume that the pseudo motions over t are linear and of a constant velocity limited within 2 pixels between the time intervals (δt). Specifically, we applied 25 motion patterns to each static light-field sample, resulting in 7,380,000 and 4,167,000 samples for 5×5 and 8×8 views, respectively. The motion patterns included linear motions in 16 directions with 2 pixels/δt, 8 directions with 1 pixel/δt, and a stationary one, as illustrated at the bottom of Fig. 4.

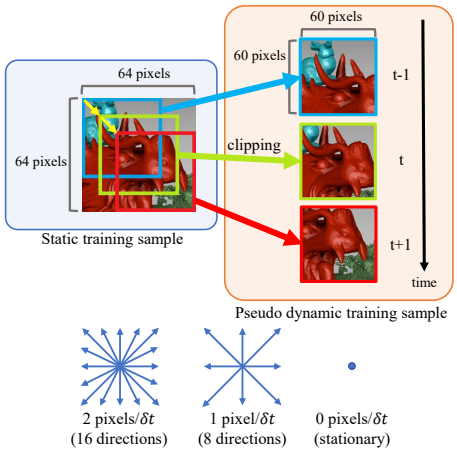

Fig. 4. Creating dynamic training samples with pseudo motions

Table 1. Datasets used for training

5×5 views (51 light fields)
Chess, Lego Bulldozer, Lego Truck, Eucalyptus Flowers, Amethyst, Bracelet, The Stanford Bunny, Jelly Beans, Lego Knights, Tarot Cards and Crystal Ball (small angular extent), Treasure Chest (Stanford [9]), Red Dragon, Happy Buddha, Messerschmitt, Dice, Green Dragon, Mini Cooper, Butterfly, Lucy (MIT [28]), Bedroom, Bicycle, Herbs, Origami, Boxes, Cotton, Sideboard, Antinous, Boardgames, Dishes, Greek, Museum, Pens, Pillows, Platonic, Rosemary, Table, Tomb, Town, Vinyl (New HCI [15]), Buddha, Buddha 2, StillLife, Papillon, MonaRoom, Medieval, Horse, Couple, Cube, Maria, Pyramide, Statue (Old HCI [14])
8×8 views (30 light fields)
Chess, Lego Bulldozer, Lego Truck, Eucalyptus Flowers, Amethyst, Bracelet, The Stanford Bunny, Jelly Beans, Lego Knights, Tarot Cards and Crystal Ball (small angular extent), Treasure Chest Bedroom (Stanford [9]), Bicycle, Herbs, Origami, Boxes, Cotton, Sideboard, Antinous, Boardgames, Dishes, Greek, Museum, Pens, Pillows, Platonic, Rosemary, Table, Tomb, Town, Vinyl (New HCI [15])

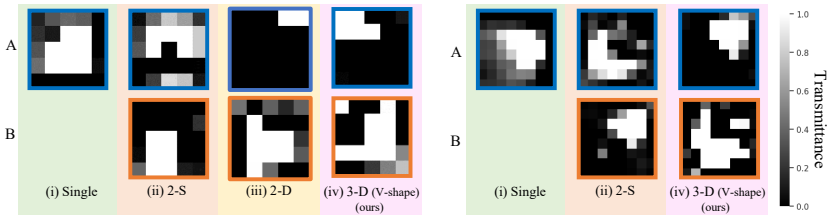

Fig. 5. Aperture patterns optimized through deep learning (left: 5×5, right: 8×8)

For each of the cases with 5×5 and 8×8 views, we used almost the same network for Single, 2-S, 2-D, and 3-D (V-shape) to make the comparison as fair as possible. The only difference was the joint between the encoder and decoder parts because the number of observed images used for reconstruction differs depending on the method (one to three). We used and extended the source code provided by Inagaki et al. [17]. The software was implemented using Python version 3.6.6 and Chainer [41] version 5.4.0. The batch size for training was set to 15. We used a built-in Adam optimizer [21]. The standard deviation of noise added to the observed image I_t was set to $\sigma = 0.005$ with respect to the image-intensity range $[0, 1]$ of I_t. The number of epochs was fixed to 20 throughout the experiments. The training with V-shape observation (our proposal) took approximately 5 days on a PC equipped with NVIDIA Geforce GTX 1080 Ti. Although the training was conducted with small image patches, the full-resolution light field could be processed at once in the testing stage because our network is fully convolutional.

The aperture patterns obtained with Single, 2-S, 2-D, and 3-D (V-shape) are shown in Fig. 5. Due to the noise added to the observed images, the resulting aperture patterns were seemingly made sufficiently bright, which helped them

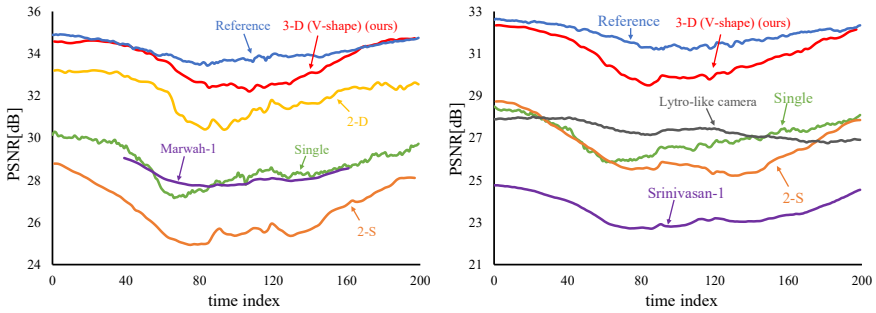

Fig. 6. Quantitative comparison of reconstructed light fields (left: 5×5, right: 8×8)

to be robust against the noise. For the case with two aperture patterns, they are optimized to be partial and complementary to each other, so that the images acquired with them should contain much disparity information with each other. Moreover, the optimization resulted in different patterns depending on the methods due to the difference of the datasets and observation patterns.

3 Experiments

We first present quantitative evaluations using a computer-generated dynamic scene. We then mention an experiment using a physical coded-aperture camera to capture a real dynamic scene.

3.1 Quantitative Evaluation

We compared the four methods denoted as Single, 2-S, 2-D, and 3-D (V-shape). Note that Single and 2-S correspond to the methods proposed by Inagaki et al. [17], which were designed exclusively for static light fields. In addition, we tested several methods that can obtain a light field from a single image at each time t: Marwah-1, Srinivasan-1, and Lytro-like camera. Marwah-1 [26] is a compressed-sensing-based method constructed on the learned dictionary and sparsity prior, where a light field with 5×5 views is reconstructed from a single coded image. Srinivasan-1 [37] is a deep-learning-based method that reconstructs 8×8 views from an ordinary image. Lytro-like camera is a simulation of a lens-array based camera [31, 32] that obtains 8×8 views simultaneously but with a less ($1/8 \times 1/8$) spatial resolution (the resulting images were upsampled to the original resolution using bicubic interpolation). We also tested a method (Reference) with which L_t is reconstructed using the same network as the one for 2-S but from two images (I_t^A and I_t^B) observed at the same t, which is practically impossible but serves as the reference that shows the upper-bound reconstruction quality.

For quantitative evaluation, we generated a light-field sequence of a dynamic scene using POV-Ray [33], which is composed of 840×593 pixels and 5×5 or

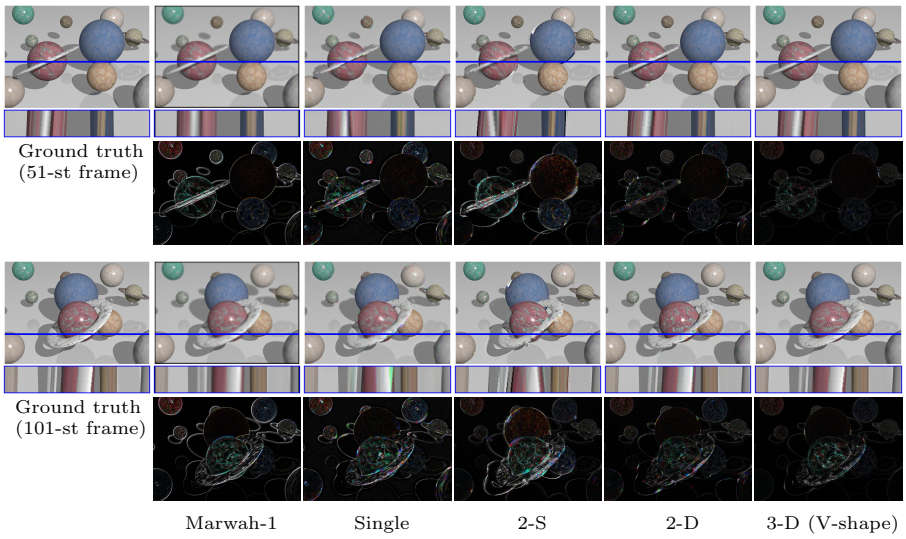

Fig. 7. Reconstructed light fields of computer-generated dynamic scene (5×5 views)

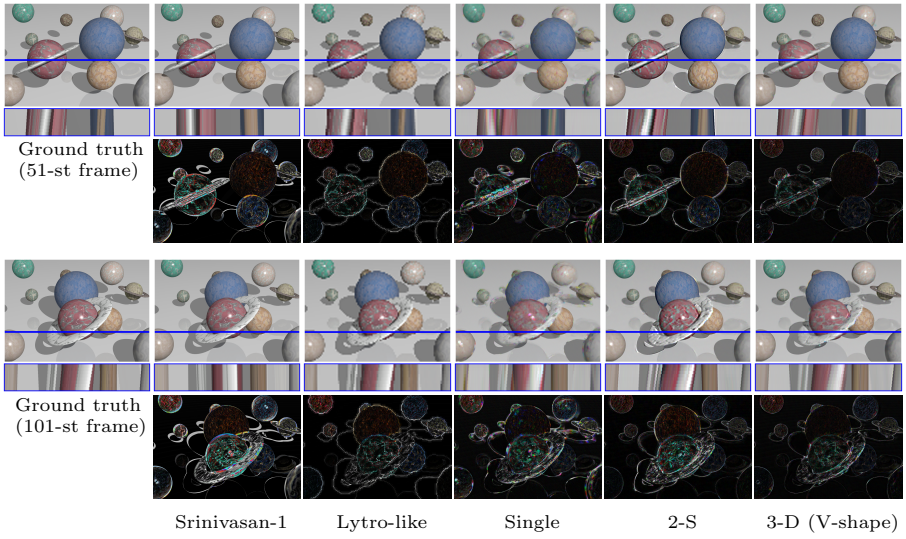

Fig. 8. Reconstructed light fields of computer-generated dynamic scene (8×8 views)

8×8 views over 200 temporal frames. This test scene contains several slowly revolving planets, producing scene motions with various velocities in various directions. At test time, we added zero-mean Gaussian noise with $\sigma = 0.005$ to the images observed from the coded aperture/mask cameras (Single, 2-S, 2-D, 3-D (V-shape), and Marwah-1) to simulate noisy imaging conditions.

Figure 6 shows the peak signal-to-noise ratios (PSNRs) (the squared errors were averaged over 25 or 64 views) over t for each method [6]. Figures 7 and 8 show the reconstructed top-left views at the 51-st and 101-st frames, for each of which an epipolar plane image (EPI) and the difference from the ground truth (magnified by 3) were also shown to better present the reconstruction quality. See the supplementary video for further details.

Several observations can be found from those results. First, reconstruction from a single observed image (Single, Marwah-1, and Srinivasan-1) was not successful because a single image alone cannot carry the disparity information. Meanwhile, Lytro-like camera can obtain correct disparities with a single shot but with the limited ($1/8 \times 1/8$) spatial resolution (zoom in on the digital version). Second, the quality of 2-S was lower than that of Single, although two observed images were provided for 2-S. In contrast, 2-D exhibited significantly better reconstruction quality than Single and 2-S. This shows the importance of the dynamic-light-field dataset over the static one to handle scene motions successfully. Finally, our proposed method (3-D (V-shape)) exhibited the best reconstruction quality among the methods, and its performance was even close to Reference. The superiority of our method over 2-D indicates the effectiveness of V-shape observation, with which the reconstruction algorithm can better separate the disparity information and scene motions. The test scene including various motions was successfully reconstructed with our method despite the fact that the dataset used for training contained only rather simple linear motions. See Appendix for more results with larger scene motions.

3.2 Experiment Using Physical Coded Aperture Camera

Finally, we acquired a dynamic light field using a physical coded aperture camera. We adopted the same hardware design as that reported in previous studies [17, 26,30,36]. The resolution of the camera (FLIR GRAS-14S5C-C) was 1384×1036 pixels, which is equivalent to the spatial resolution of light field acquired with it. We used a Nikon Rayfact lens (25 mm F1.4 SF2514MC). The aperture was implemented using an liquid crystal on silicon (LCoS) display (Forth Dimension Displays, SXGA-3DM) with 1280×1024 pixels. We divided the central area (750×750 pixels) of the LCoS display into 5×5 regions (each with 150×150 pixels), which corresponded to the angular resolution (the number of views) of 5×5. The exposure time was set to 40 msec. Due to the hardware constraint, the frame rate for the observed images was approximately 12 frames per second, and the light field video was reconstructed with 6 frames per second.

The experimental setup and camera we used are illustrated in Fig. 9 (top-left and top-center). The target scene consisted of three objects located on a motorized turntable, which produced various scene motions. We used two sets of aperture patterns corresponding to 2-S and 3-D (V-shape). The reconstruction was carried out with the respective reconstruction networks. Some examples

[6] For Marwah-1, only some of the frames were reconstructed due to heavy computation. It took approximately 20 hours to reconstruct a single light field.

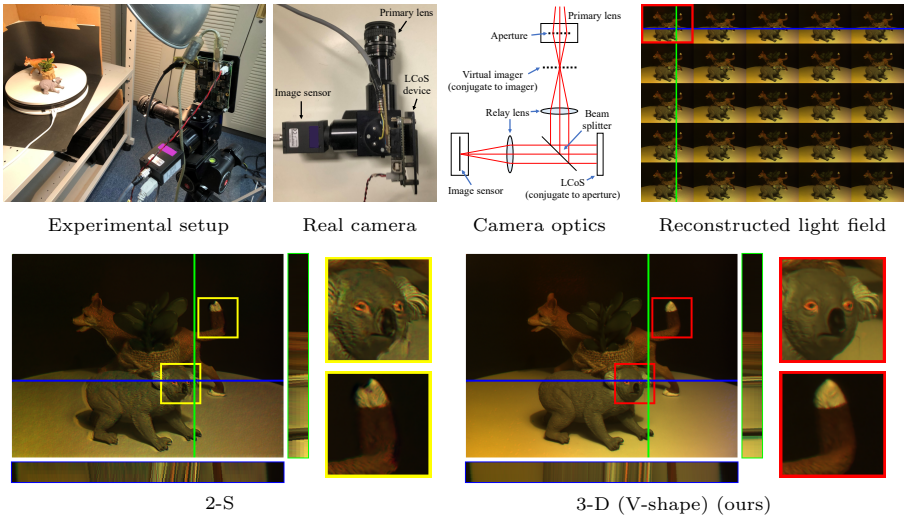

Fig. 9. Experiment using physical coded-aperture camera (see supplementary video)

of the reconstructed top-left views and EPIs are shown in Fig. 9 (bottom). The result from 2-S seems quite poor; unnatural object edges and incorrect disparities are noticeable. Our method exhibited fine reconstruction quality over all the viewpoints over t. See the supplementary video for more details.

4 Conclusions

We developed a method of acquiring a dynamic light field through a coded aperture camera, where the entire process of light field acquisition (including the aperture patterns for the camera) is modeled as a deep CNN and optimized through training on a large amount of light-field data. Our contribution is twofold, both of which are indispensable to successfully handle a dynamic light field. We first introduced a new configuration of image observation called V-shape observation to help the CNN successfully separate the disparity information from scene motions. We then constructed a dynamic light field dataset (constructed from the static dataset with pseudo motions) to train the CNN, which makes it more adaptable to dynamic scenes. To our knowledge, this is the first work that achieves compressive acquisition of a dynamic light field based on the concept of deep optics, which will inspire further development of computational cameras. Our future work will include extending the training dataset to cover a larger amount of motions and exploring better network structures and color-processing methods. Exploring other input configurations (with different numbers of aperture patterns and observed images) would also be an interesting direction.

References

1. Adelson, E.H., Bergen, J.R.: The plenoptic function and the elements of early vision. In: Computational Models of Visual Processing. pp. 3–20 (1991)
2. Adelson, E.H., Wang, J.Y.: Single lens stereo with a plenoptic camera. IEEE transactions on pattern analysis and machine intelligence **14**(2), 99–106 (1992)
3. Arai, J., Okano, F., Hoshino, H., Yuyama, I.: Gradient-index lens-array method based on real-time integral photography for three-dimensional images. Applied optics **37**(11), 2034–2045 (1998)
4. Babacan, S.D., Ansorge, R., Luessi, M., Mataran, P.R., Molina, R., Katsaggelos, A.K.: Compressive light field sensing. IEEE Transactions on image processing **21**(12), 4746–4757 (2012)
5. Bishop, T.E., Zanetti, S., Favaro, P.: Light field superresolution. In: Computational Photography (ICCP), 2009 IEEE International Conference on. pp. 1–9. IEEE (2009)
6. Candes, E.J., Eldar, Y.C., Needell, D., Randall, P.: Compressed sensing with coherent and redundant dictionaries. Applied and Computational Harmonic Analysis **31**(1), 59–73 (2011)
7. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE signal processing magazine **25**(2), 21–30 (2008)
8. Chen, B., Ruan, L., Lam, M.L.: LFGAN: 4d light field synthesis from a single rgb image. ACM Trans. Multimedia Comput. Commun. Appl. **16**(1) (Feb 2020)
9. Computer Graphics Laboratory, Stanford University: The (new) stanford light field archive (2018), `http://lightfield.stanford.edu`
10. Donoho, D.L.: Compressed sensing. IEEE Transactions on information theory **52**(4), 1289–1306 (2006)
11. Fujii, T., Mori, K., Takeda, K., Mase, K., Tanimoto, M., Suenaga, Y.: Multipoint measuring system for video and sound-100-camera and microphone system. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 437–440. IEEE (2006)
12. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. pp. 43–54 (1996)
13. Gupta, M., Jauhari, A., Kulkarni, K., Jayasuriya, S., Molnar, A., Turaga, P.: Compressive light field reconstructions using deep learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1277–1286 (2017)
14. Heidelberg Collaboratory for Image Processing: Datasets and benchmarks for densely sampled 4D light fields. `http://lightfieldgroup.iwr.uni-heidelberg.de/?page\_id=713` (2016)
15. Heidelberg Collaboratory for Image Processing: 4D light field dataset (2018), `http://hci-lightfield.iwr.uni-heidelberg.de/`
16. Huang, F.C., Chen, K., Wetzstein, G.: The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. ACM Transactions on Graphics (TOG) **34**(4), 60 (2015)
17. Inagaki, Y., Kobayashi, Y., Takahashi, K., Fujii, T., Nagahara, H.: Learning to capture light fields through a coded aperture camera. In: The European Conference on Computer Vision (ECCV) (September 2018)
18. Isaksen, A., McMillan, L., Gortler, S.J.: Dynamically reparameterized light fields. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. pp. 297–306 (2000)

19. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016) **35**(6) (2016)
20. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1646–1654 (2016)
21. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: The International Conference on Learning Representations (ICLR) (2015)
22. Lee, S., Jang, C., Moon, S., Cho, J., Lee, B.: Additive light field displays: realization of augmented reality with holographic optical elements. ACM Transactions on Graphics (TOG) **35**(4), 60 (2016)
23. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 31–42. ACM (1996)
24. Liang, C.K., Lin, T.H., Wong, B.Y., Liu, C., Chen, H.H.: Programmable aperture photography: multiplexed light field acquisition. ACM Transactions on Graphics (TOG) **27**(3), 55 (2008)
25. Maeno, K., Nagahara, H., Shimada, A., Taniguchi, R.I.: Light field distortion feature for transparent object recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2786–2793 (2013)
26. Marwah, K., Wetzstein, G., Bando, Y., Raskar, R.: Compressive light field photography using overcomplete dictionaries and optimized projections. ACM Transactions on Graphics (TOG) **32**(4), 46 (2013)
27. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) (2019)
28. MIT Media Lab's Camera Culture Group: Compressive light field camera, `http://cameraculture.media.mit.edu/projects/compressive-light-field-camera/`
29. Nabati, O., Mendlovic, D., Giryes, R.: Fast and accurate reconstruction of compressed color light field. In: 2018 IEEE International Conference on Computational Photography (ICCP). pp. 1–11 (May 2018)
30. Nagahara, H., Zhou, C., Watanabe, T., Ishiguro, H., Nayar, S.K.: Programmable aperture camera using LCoS. In: European Conference on Computer Vision. pp. 337–350. Springer (2010)
31. Ng, R.: Digital light field photography. Ph.D. thesis, stanford university (2006)
32. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Computer Science Technical Report CSTR **2**(11), 1–11 (2005)
33. Persistence of Vision Pty. Ltd.: Persistence of vision raytracer (version 3.6) (2004), `http://www.povray.org/`
34. Shi, L., Hassanieh, H., Davis, A., Katabi, D., Durand, F.: Light field reconstruction using sparsity in the continuous fourier domain. ACM Transactions on Graphics (TOG) **34**(1), 12 (2014)
35. Shin, C., Jeon, H.G., Yoon, Y., Kweon, I.S., Kim, S.J.: EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images. In: IEEE CVPR. pp. 4748–4757 (2018)
36. Sonoda, T., Nagahara, H., Taniguchi, R.: Motion-invariant coding using a programmable aperture camera. IPSJ Transactions on Computer Vision and Applications **6**, 25–33 (6 2014)

37. Srinivasan, P.P., Wang, T., Sreelal, A., Ramamoorthi, R., Ng, R.: Learning to synthesize a 4D RGBD light field from a single image. In: IEEE International Conference on Computer Vision. pp. 2262–2270 (2017)
38. Taguchi, Y., Koike, T., Takahashi, K., Naemura, T.: TransCAIP: A live 3D TV system using a camera array and an integral photography display with interactive control of viewing parameters. IEEE Transactions on Visualization and Computer Graphics **15**(5), 841–852 (Sept 2009)
39. Takahashi, K., Kobayashi, Y., Fujii, T.: From focal stack to tensor light-field display. IEEE Transactions on Image Processing **27**(9), 4571–4584 (Sep 2018)
40. Tambe, S., Veeraraghavan, A., Agrawal, A.: Towards motion aware light field video for dynamic scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1009–1016 (2013)
41. Tokui, S., Oono, K., Hido, S., Clayton, J.: Chainer: a next-ngeneration open source framework for deep learning. In: Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS) (2015)
42. Vadathya, A.K., Girish, S., Mitra, K.: A unified learning based framework for light field reconstruction from coded projections. IEEE Transactions on Computational Imaging pp. 1–1 (2019)
43. Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., Tumblin, J.: Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. ACM Transactions on Graphics (TOG) **26**(3), 69 (2007)
44. Wang, T.C., Efros, A.A., Ramamoorthi, R.: Depth estimation with occlusion modeling using light-field cameras. IEEE transactions on pattern analysis and machine intelligence **38**(11), 2170–2181 (2016)
45. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A., Ramamoorthi, R.: A 4d light-field dataset and cnn architectures for material recognition. In: European Conference on Computer Vision (ECCV). pp. 121–138 (2016)
46. Wang, T.C., Zhu, J.Y., Kalantari, N.K., Efros, A.A., Ramamoorthi, R.: Light field video capture using a learning-based hybrid imaging system. ACM Trans. Graph. **36**(4), 133:1–133:13 (2017)
47. Wang, Y., Liu, F., Wang, Z., Hou, G., Sun, Z., Tan, T.: End-to-end view synthesis for light field imaging with pseudo 4dcnn. In: The European Conference on Computer Vision (ECCV) (September 2018)
48. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. IEEE transactions on pattern analysis and machine intelligence **36**(3), 606–619 (2014)
49. Wetzstein, G., Lanman, D., Hirsch, M., Raskar, R.: Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting. ACM Trans. Graph. (Proc. SIGGRAPH) **31**(4), 1–11 (2012)
50. Wilburn, B., Joshi, N., Vaish, V., Talvala, E.V., Antunez, E., Barth, A., Adams, A., Horowitz, M., Levoy, M.: High performance imaging using large camera arrays. ACM Transactions on Graphics (TOG) **24**(3), 765–776 (2005)
51. Williem, W., Park, I.K., Lee, K.M.: Robust light field depth estimation using occlusion-noise aware data costs. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2017)
52. Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field reconstruction using deep convolutional network on EPI. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1638–1646 (July 2017)

53. Yagi, Y., Takahashi, K., Fujii, T., Sonoda, T., Nagahara, H.: PCA-coded aperture for light field photography. In: IEEE International Conference on Image Processing (ICIP) (2017)
54. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. In: SIGGRAPH (2018)