PAPER Special Section on Picture Coding and Image Media Processing

Time-Multiplexed Coded Aperture and Coded Focal Stack –Comparative Study on Snapshot Compressive Light Field Imaging

Kohei TATEISHI[†a)], **Chihiro TSUTAKE**[†], ***Members***, **Keita TAKAHASHI**[†], ***Senior Member***, ***and*** **Toshiaki FUJII**[†], ***Fellow***

SUMMARY A light field (LF), which is represented as a set of dense, multi-view images, has been used in various 3D applications. To make LF acquisition more efficient, researchers have investigated compressive sensing methods by incorporating certain coding functionalities into a camera. In this paper, we focus on a challenging case called snapshot compressive LF imaging, in which an entire LF is reconstructed from only a single acquired image. To embed a large amount of LF information in a single image, we consider two promising methods based on rapid optical control during a single exposure: time-multiplexed coded aperture (TMCA) and coded focal stack (CFS), which were proposed individually in previous works. Both TMCA and CFS can be interpreted in a unified manner as extensions of the coded aperture (CA) and focal stack (FS) methods, respectively. By developing a unified algorithm pipeline for TMCA and CFS, based on deep neural networks, we evaluated their performance with respect to other possible imaging methods. We found that both TMCA and CFS can achieve better reconstruction quality than the other snapshot methods, and they also perform reasonably well compared to methods using multiple acquired images. To our knowledge, we are the first to present an overall discussion of TMCA and CFS and to compare and validate their effectiveness in the context of compressive LF imaging.

key words: *light field, compressive sensing, coded aperture, focal stack*

1. Introduction

A light field (LF) [1], [2] is represented as a set of images taken from many (dozens) of viewpoints aligned regularly at small intervals. LFs are used in many 3D applications such as synthetic refocusing [3], [4], view synthesis [5], depth estimation [6], [7], and 3D displays [8], [9].

Acquisition of a dense LF is a challenging task because of the large amount of data involved. Several researchers have applied a direct approach using multiple cameras [10], [11] or a camera mounted on a moving gantry [12], which is costly in terms of the hardware or the time required to capture an entire LF. To make the acquisition process more efficient, lens-array-based cameras [4] and coded aperture (CA) cameras [13]–[16] have been investigated. Lens-array-based cameras enable single-shot acquisition of an entire LF at the cost of the spatial resolution for each viewpoint: in principle, $U \times V$ views are obtained with the $1/U \times 1/V$ spatial resolution of the image sensor. In contrast, with a CA camera, an entire LF with the full sensor spatial resolution can be computationally reconstructed from observed images. Moreover, the CA method enables compressive imaging, because the number of coded images used for reconstruction is typically two to four, which is much less than the number of viewpoints in the target LF. A focal stack (FS), which is a set of differently focused images, can also be used for compressive light field imaging [17], [18], because only a few images taken with different focus depths are sufficient to computationally reconstruct an entire LF.

In this paper, we focus on compressive imaging of an LF by using a CA and an FS. In particular, we are interested in snapshot compressive imaging, an extreme case of compressive imaging in which an entire LF is obtained from only a single observed image. In most previous methods [13]–[18], however, multiple images taken with different coding patterns or focus depths were necessary for high-quality LF reconstruction. A possible solution to overcome this limitation is to exploit rapid optical control during a single exposure. Specifically, we consider two promising methods with this kind of rapidly controlled imaging: time-multiplexed coded aperture (TMCA) [19], [20] and coded focal stack (CFS) [21], in which the CA and FS methods, respectively, are combined with pixel-wise exposure coding within a single exposure.

A key point is that both TMCA and CFS can be interpreted in a unified manner: these methods can be regarded as embedding multiple coded/focused images into a single observed image through pixel-wise exposure coding. Hence, we have developed a unified algorithm pipeline for TMCA and CFS. Specifically, we model the entire imaging pipeline for TMCA and CFS via deep neural networks, and we jointly optimize the coding patterns for image acquisition and the algorithm for LF reconstruction from the acquired image. Moreover, we have experimentally compared the performance of TMCA and CFS with that of other possible methods for compressive LF imaging. As a result, we found that both TMCA and CFS can achieve better reconstruction quality than the other snapshot methods, and that they perform reasonably well in comparison to the use of multiple images acquired with the CA and FS methods.

Note that TMCA and CFS were proposed individually in previous works (TMCA: [19], [20]; CFS: [21]), and CFS was not directly used for compressive LF imaging. To our knowledge, we are the first to present an overall discussion of TMCA and CFS and to compare and validate their effectiveness in the context of compressive LF imaging. We believe that our work will contribute to extension of the frame-

Manuscript received January 4, 2022.
Manuscript revised March 15, 2022.
Manuscript publicized May 26, 2022.

[†]The authors are with the Graduate School of Engineering, Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: tateishi@fujii.nuee.nagoya-u.ac.jp

DOI: 10.1587/transinf.2022PCP0003

work of compressive LF imaging to cover various imaging methods based on different camera architectures.

2. Imaging Models for TMCA and CFS

2.1 Configuration

As shown in Fig. 1, we parameterize a light ray traveling inside a camera with four variables (x, y, u, v), where (u, v) and (x, y) denote the intersections of the light ray with the aperture and imaging planes, respectively. We assume the coordinate (x, y, u, v) to be discretized. The LF is defined as a 4D function $l(x, y, u, v)$ that returns the light intensity for a given 4D coordinate. Note that $l(x, y, u, v)$ also represents a set of multi-view images, where (u, v) and (x, y) denote the viewpoint and the pixel position, respectively. In other words, (u, v) and (x, y) correspond to the respective angular and spatial dimensions. For simplicity, we assume that an LF has only one color channel. For an LF with RGB colors, we treat each of the channels individually. We also assume that the target scene is stationary, and thus, that the LF $l(x, y, u, v)$ does not vary during the image acquisition process. Our goal is to obtain the entire LF $l(x, y, u, v)$ from the images taken by the camera.

In a standard camera, all the light rays that reach the same pixel (x, y) are added together. The resulting image $i(x, y)$ is given as

$$i(x, y) = \sum_{u,v} l(x, y, u, v), \tag{1}$$

where the information along the angular dimension $((u, v))$ is mostly lost and hard to recover. To better embed the original information of $l(x, y, u, v)$ into $i(x, y)$, a mechanism to modulate the light rays is necessary. In the remainder of this section, we describe several image acquisition methods for compressive LF imaging, where the target LF can be computationally reconstructed from a smaller number of observed images. In particular, we introduce two methods for snapshot compressive LF imaging: TMCA and CFS.

2.2 Coded Aperture to Time-Multiplexed Coded Aperture

A popular choice for compressive LF imaging is coded aperture (**CA**) imaging [13]–[16], in which a light-attenuating mask is inserted at the aperture plane to encode the light rays. Through modulation by a light-attenuating mask pattern $a(u, v) \in [0, 1]$, the image $i(x, y)$ is given as

$$i(x, y) = \sum_{u,v} a(u, v) l(x, y, u, v). \tag{2}$$

A single modulation pattern for the angular dimension, $a(u, v)$, is applied to each obtained image, which limits the capability of information embedding. Therefore, multiple images (taken with different coding patterns) are usually used to obtain a high-quality LF.

To enhance the coding capability of the CA method, time-multiplexed coded aperture (**TMCA**) [19], [20] (called factorized modulation in [19]) was introduced. TMCA can be implemented by combining CA and pixel-wise exposure coding [22], [23], which are synchronously varied during an exposure. We assume that the exposure time is divided into T discrete time slots, and T sets of coding patterns are applied during the exposure. Specifically, the image $i(x, y)$ is given as

$$i(x, y) = \frac{1}{T} \sum_{t=0}^{T-1} \sum_{u,v} p_t(x, y) a_t(u, v) l(x, y, u, v), \tag{3}$$

where $a_t(u, v)$ is a time-varying CA pattern, and $p_t(x, y)$ is a time-varying pixel-wise exposure pattern (on/off) on the imaging plane. If $p_t(x, y)$ is disabled (i.e., $p_t(x, y) = 1$), then Eq. (3) reduces to Eq. (2), where $a(u, v)$ is the average of $a_t(u, v)$ over time. Through this approach, TMCA is effective for snapshot compressive LF imaging, in which an entire LF is obtained from only a single observed image $i(x, y)$.

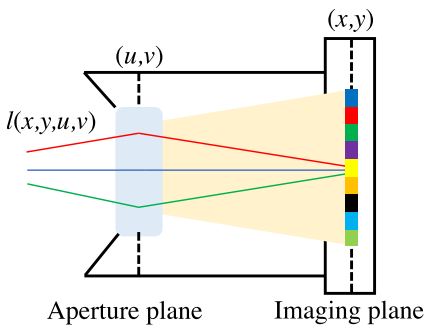


Fig. 1 Camera and light field.

2.3 Focal Stack to Coded Focal Stack

A focal stack (**FS**), which is a set of differently focused images, contains 3D information for the target scene. An entire LF can be reconstructed from a focal stack consisting of only a few images [17], [18]. Consequently, a focal stack can also be used for compressive LF imaging[†].

An image focused at a specific depth is modeled by a shear-and-add operation on the LF as follows:

$$f_d(x, y) = \sum_{u,v} l(x - d(u - u_c), y - d(v - v_c), u, v), \tag{4}$$

where (u_c, v_c) is the central viewpoint of the LF, and d is a shear parameter corresponding to the focus depth. However, multiple images with different focus depths are necessary for high-quality LF reconstruction.

Lin et al. [21] proposed a compressive imaging method for an FS called a coded focal stack (**CFS**), in which multiple focused images are embedded into a single observed

[†]It was shown in [24], [25] that arbitrary views can be analytically computed from a sufficiently dense focal stack that consists of dozens of images focused at slightly different depths. This means that such a dense focal stack is almost equivalent to the corresponding LF. In contrast, we are interested in much sparser focal stacks for the purpose of compressive LF imaging.

image through pixel-wise exposure coding during an exposure. Although they did not apply CFS directly for LF reconstruction, we hypothesize that it can be used for snapshot compressive LF imaging.

Here, we formulate CFS to account for the fact that the focal position can continuously change during an exposure [26]. Let the range of the exposure time be $[0, 1]$; then, the focus depth $d(\tau)$ at time $\tau \in [0, 1]$ is defined as

$$d(\tau) = -D_{\max} + 2D_{\max}\tau, \tag{5}$$

where the range for $d(\tau)$ is defined as $[-D_{\max}, D_{\max}]$. The image obtained with CFS is given as

$$i(x, y) = \int_0^1 p_\tau(x, y) f_{d(\tau)}(x, y) d\tau. \tag{6}$$

This equation can be interpreted as indicating that multiple images focused at different depths $f_{d(\tau)}(x, y)$ are encoded by $p_\tau(x, y)$ and fused into a single observed image $i(x, y)$. We assume that the pixel-wise exposure pattern $p_\tau(x, y)$ is controlled discretely over time τ, and we denote the t-th pattern by $p_t(x, y)$ $(t = 0, \ldots, T-1)$:

$$p_\tau(x, y) = p_t(x, y), \quad \tau \in \left[\frac{t}{T}, \frac{t+1}{T}\right]. \tag{7}$$

By substituting Eq. (7) into Eq. (6), we obtain

$$i(x, y) = \frac{1}{T} \sum_{t=0}^{T-1} p_t(x, y) \int_{t/T}^{(t+1)/T} f_{d(\tau)}(x, y) d\tau. \tag{8}$$

We thus expect that the original LF $l(x, y, u, v)$ can be reconstructed from only a single observed image $i(x, y)$.

2.4 Summary

We conclude this section by describing the similarity between TMCA and CFS. Both TMCA and CFS can be interpreted in a unified manner as extensions of the CA and FS methods, respectively. Specifically, Eqs. (3) and (8) can be represented in the same form as

$$i(x, y) = \frac{1}{T} \sum_{t=0}^{T-1} p_t(x, y) j_t(x, y), \tag{9}$$

where $j_t(x, y)$ is given by Eq. (10) or (11) for TMCA or CFS, respectively:

$$j_t(x, y) = \sum_{u,v} a_t(u, v) l(x, y, u, v), \tag{10}$$

$$j_t(x, y) = \int_{t/T}^{(t+1)/T} f_{d(\tau)}(x, y) d\tau. \tag{11}$$

In both cases, the image acquisition process is divided into two steps: the target LF $l(x, y, u, v)$ is first compressed into T images $j_t(x, y)$ by using either the CA or FS method; then, these images are further fused into a single observed image $i(x, y)$ through the pixel-wise exposure coding $p_t(x, y)$. Our goal is to reconstruct the original LF $l(x, y, u, v)$ from only the single observed image $i(x, y)$.

3. Algorithm Pipeline

Here, we introduce an algorithm pipeline that can handle both TMCA and CFS in a unified manner, so that an entire LF can be obtained from a single observed image.

In the remainder of this paper, we assume that the target LF has 5×5 viewpoints and $W \times H$ pixels. Accordingly, the aperture coding pattern $a_t(u, v)$ has $5 \times 5 \times T$ elements. Considering hardware restrictions [23], we also assume that the pixel-wise exposure coding $p_t(x, y)$ takes a repeating pattern with a cycle of 8×8 pixels. Therefore, $p_t(x, y)$ effectively has only $8 \times 8 \times T$ elements.

The entire pipeline is illustrated in Fig. 2. In either TMCA or CFS, a target LF (with $5 \times 5 \times W \times H$ elements) is compressed into an image (with $W \times H$ elements) through the image acquisition process. Then, computational reconstruction is performed to recover the original LF (with $5 \times 5 \times W \times H$ elements) from the image. We jointly optimize the image acquisition process ($a_t(u, v)$ and $p_t(x, y)$) and the LF reconstruction process in a deep-learning framework. To this end, we implemented these processes by using PyTorch, a Python-based framework for deep neural networks. The entire pipeline is trained end to end so that the difference between the input LF and output LF is minimized. This kind of deep-learning-based approach has proven to be quite successful for compressive LF imaging tasks [14], [20], [27].

The image acquisition process was implemented using trainable parameters, which were optimized in the training process by using the Autograd functionality in PyTorch. For TMCA, we take all the elements of $a_t(u, v)$ and $p_t(x, y)$ as trainable parameters, and we implement the computation process of Eq. (3). For CFS, we first compute T images by using Eq. (11), where $\tau \in [0, 1]$ is evenly discretized to 16 levels and $D_{\max}$ is set to 2 (unless specified otherwise) in Eq. (5). We then take the elements of $p_t(x, y)$ as trainable parameters and perform the computation of Eq. (9). For either TMCA or CFS, the target LF is finally compressed into a single observed image $i(x, y)$ through the image acquisition process. To account for noise during the imaging process, we contaminate $i(x, y)$ with additive zero-mean Gaussian noise having $\sigma = 0.005$ w.r.t. the intensity range of $[0, 1]$ for $i(x, y)$.

The LF reconstruction process was implemented as a deep convolutional neural network (CNN). The input to the reconstruction process is a single observed image $i(x, y)$. Because $i(x, y)$ is encoded with an 8×8-pixel repetitive pattern $p_t(x, y)$, we represent the observed image $i(x, y)$ as a set of 64 (8×8) sub-sampled images and feed it to the reconstruction network. By doing so, we can clarify the pixel groups encoded by the same exposure patterns over time; otherwise, the network would not be informed of the structure of $i(x, y)$. Specifically, we rewrite $p_t(x, y)$ as

$$p_t(x, y) = q_t(k, l), \quad (k, l) = (x\%8, y\%8), \tag{12}$$

where % is the modulo operator. By using this relation,

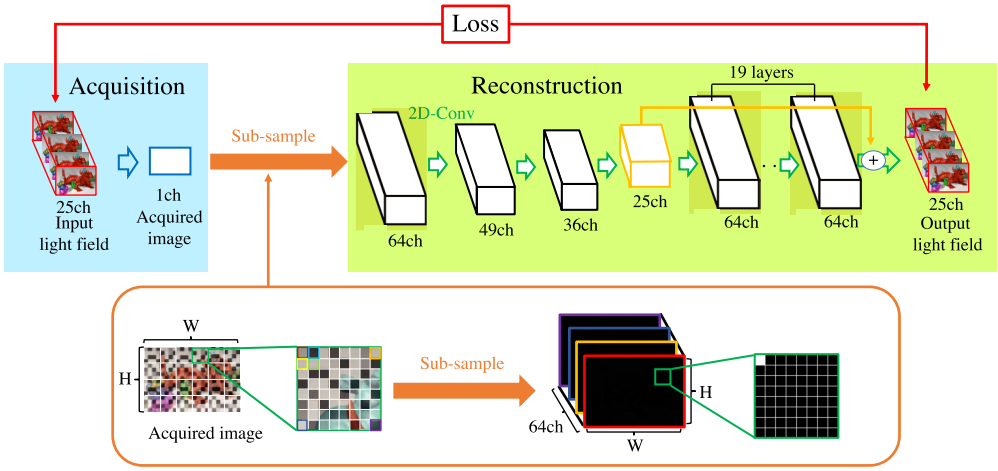


Fig. 2 Network architecture.

Eq. (9) can be written as

$$i(8x'+k,8y'+l)=\frac{1}{T}\sum_{t=0}^{T-1}q_t(k,l)j_t(8x'+k,8y'+l), \quad (13)$$

where $(x,y)=(8x'+k,8y'+l)$. In accordance with this structure, we define each sub-sampled image $i_{k,l}(x,y)$ $(k,l\in\{0,1,\ldots,7\})$ as

$$i_{k,l}(x,y)=\begin{cases} i(x,y) & (x\%8,y\%8)=(k,l) \\ 0 & \text{otherwise} \end{cases}, \quad (14)$$

where each $i_{k,l}(x,y)$ includes only the pixels that are encoded with $q_t(k,l)$. All the sub-sampled images are stacked along the channel dimension and fed to the reconstruction network.

The reconstruction network was designed as a stack of 2D convolutional layers with a residual connection. The kernel size is 5×5 for the first three convolutional layers, and 3×3 for the other layers. Each convolutional layer except for the first three and the final ones is followed by rectified linear unit (ReLU) activation. The spatial size of the processed data is kept unchanged throughout the network. The output has 25 channels, which correspond to 25 viewpoints of the reconstructed LF. Note that what we describe here is merely one of the feasible architectures to implement the task of LF reconstruction. We adopted a rather simple architecture, similar to that of Inagaki et al. [14], to balance the reconstruction quality and computational efficiency. Further quality improvement would be expected by replacing this architecture with a better one. However, our main focus is not the network architecture but comparison and evaluation of different imaging methods for compressive LF imaging. Accordingly, we kept the network architecture simple and flexible enough to accommodate various imaging methods.

As mentioned above, we jointly train the image acquisition and LF reconstruction processes end to end. We use the mean squared error (MSE) between the target and reconstructed LFs as the loss function. We use the Adam optimizer to control the learning rate. In each parameter update step, we clip the elements of $a_t(u,v)$ into the range of [0,1], because $a_t(u,v)$ should take transmittance values. We also binarize the elements of $p_t(x,y)$, because they should take on/off values. To control this binarization, we specify the exposure ratio R, such that the top $100\times R$ % of the elements are set to 1, while the others are set to 0. R is set to 0.75 unless specified otherwise.

For the experiments reported here, we followed Inagaki et al. [14] in preparing the training dataset. We collected training samples from 51 LFs found in public datasets [28]–[31]. Each sample consisted of 5×5 views with 64×64 pixels. Since our method assumed the input/output LFs to be monochrome, the RGB channels of each LF were treated as three monochrome LFs; we made no distinction with respect to the color. We also applied six-level intensity augmentation. The number of training samples was 295,200. We trained the entire networks (including both the image acquisition and LF reconstruction processes) for TMCA and CFS over 20 epochs, which took 8.5 and 22 hours, respectively, on an NVIDIA GeForce RTX 2080 Ti GPU. Meanwhile, the LF reconstruction process on our pre-trained networks took 160 ms for a target LF with 512×512 pixels. More precisely, the reported computation time includes three iterations of the reconstruction process for the three color components of a target LF. We used the same pre-trained network for the three color components; thus, the set of coding patterns and reconstruction process were identical for the three color components.

4. Experiments

For evaluation we used four LFs—"Dino," "Kitchen," "Medieval2," and "Tower"—taken from a public dataset [31]. These LFs were not included in our training dataset. Each LF has 5×5 views with 512×512 pixels. The reconstruction quality was evaluated in terms of the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM). The PSNR values were calculated from the MSEs over all the pixels, viewpoints, and color channels. The SSIM val-

ues were averaged over all the viewpoints. These quality metrics were computed for the central 510 × 510 pixels of the reconstructed LFs, because a Lytro-like method cannot reconstruct peripheral pixels.

4.1 Evaluation of TMCA and CFS

Both TMCA and CFS have several preset parameters, and we analyzed the effects of the parameters to characterize these methods. The results are summarized in Fig. 3 for various values of T, the number of patterns applied within an exposure (see Eq. (9)); R, the ratio of exposed pixels in the imaging plane (see Sect. 3); and $D_{\max}$, the range of focus depths for CFS (see Eq. (5)). Here, we changed only one parameter at a time from the default values ($T = 2$, $R = 0.75$, and $D_{\max} = 2$) that we ultimately chose. As seen in (a), the reconstruction quality improved significantly as T increased from 1 to 2. However, there was no clear advantage to further increasing the value of T, and we thus chose $T = 2$. Similarly, the results seen in (b) and (c) respectively indicate that $R = 0.75$ was a reasonable choice for both TMCA and CFS, and that $D_{\max} = 2$ was a suitable choice for the target LFs.

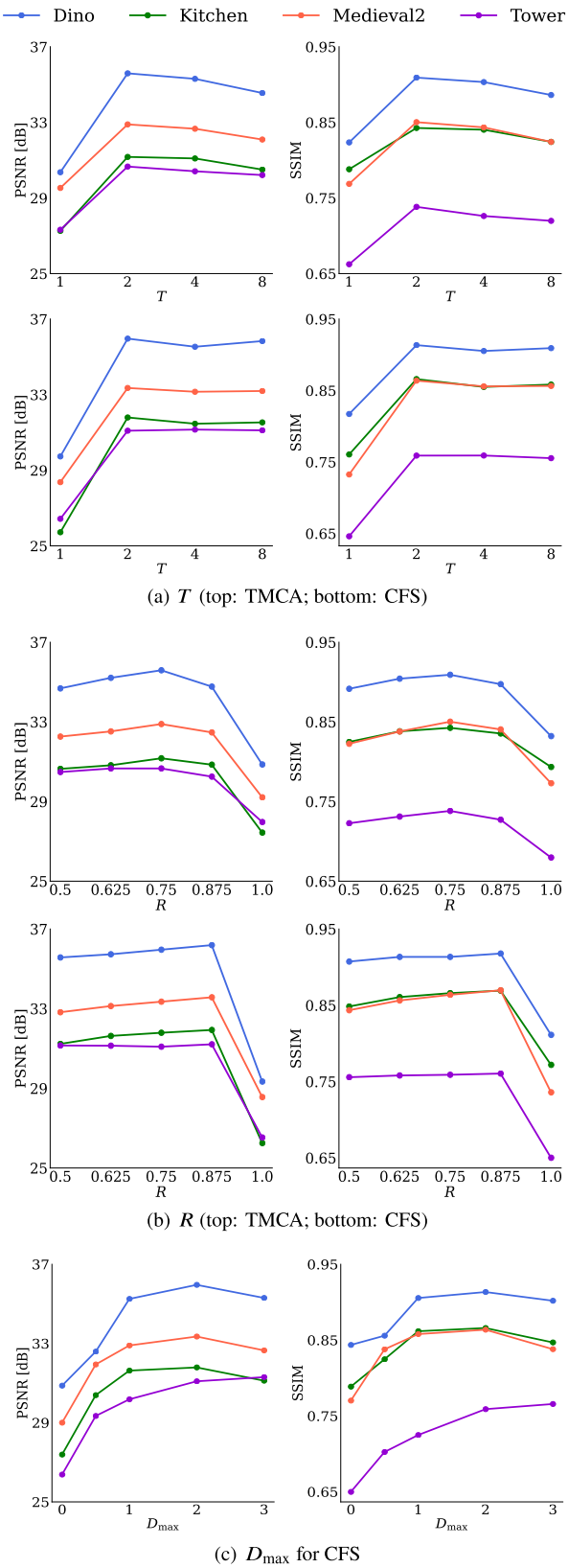
We present an interpretation for the performance against T shown in Fig. 3(a). As mentioned in Sect. 2.4, the imaging processes of TMCA and CFS can be regarded as a two-steps compression; 25 images in a light field are first compressed into T images, and then, these T images are further reduced into a single image. A larger T can help to preserve more information in the first step, but it will cause a greater information loss in the second step. Due to this trade-off between the first and second steps with respect to the preserved/lost information, increasing T above 2 did not lead to obvious quality improvement.

We also analyzed the effectiveness of the sub-sampled image structure represented by Eq. (14). Specifically, we tested an ablation case without sub-sampling, in which the acquired image $i(x, y)$ was fed directly to the reconstruction network (the network with this modification was trained from scratch). As shown in Fig. 4, the sub-sampled image structure yielded a higher reconstruction quality than the ablation case. This result supports our assumption that the periodic structure of pixel-wise exposure coding should be informed to the network.

Figure 5 shows the coding patterns for TMCA and CFS that were obtained with the default parameters ($T = 2$, $R = 0.75$, and $D_{\max} = 2$). Note that these patterns were optimized with the reconstruction network by the end-to-end training process performed on the training dataset. Although intuitive interpretations of these patterns are not straightforward, we can see that for each method the two temporal patterns supplement each other, which should make acquisition of the LF information more effective.

4.2 Comparison with Other Imaging Methods

We also evaluated the performance of TMCA and CFS in



(a) T (top: TMCA; bottom: CFS)

(b) R (top: TMCA; bottom: CFS)

(c) $D_{\max}$ for CFS

Fig. 3 Performance analysis with respect to the preset parameters. (a) T: the number of patterns applied within an exposure. (b) R: the ratio of exposed pixels in the imaging plane. (c) $D_{\max}$: the range of depths for CFS. In each row, the PSNR (left) and SSIM (right) scores are shown for four LFs.

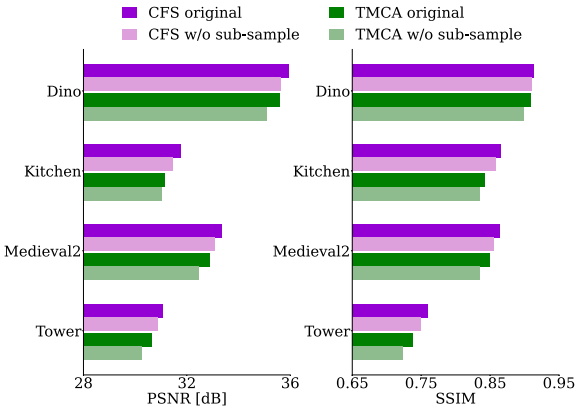


Fig. 4 Results of an ablation study for the sub-sampled image structure, showing the PSNR (left) and SSIM (right) scores for four LFs.

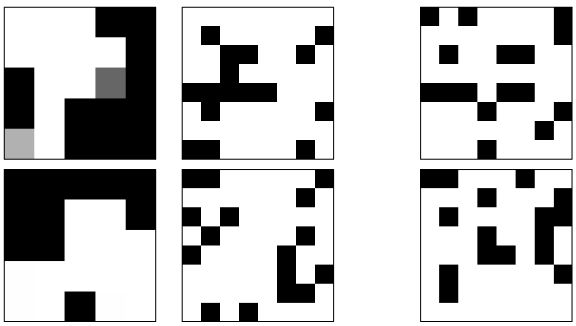


Fig. 5 Coding patterns for TMCA (left) and CFS (right) with T=2.

comparison with other methods for compressive LF imaging. To make the comparisons as fair as possible, we implemented those methods with minimal changes from TMCA and CFS. Although the imaging process differs from method to method, we kept the LF reconstruction process unchanged except for the input to the reconstruction network. We trained the other methods on the same training dataset for the same number of epochs as we did for TMCA and CFS.

4.2.1 CA and FS

Both CA and FS were considered as targets for comparison, as TMCA and CFS derive from CA and FS, respectively. For CA, we implemented the process of acquiring $N \in \{1, 2, 3\}$ images by using Eq. (2) with a distinct coding pattern $a(u, v)$ for each acquisition. This involved $5 \times 5 \times N$ trainable parameters. For FS, we implemented the process of acquiring $N \in \{1, 2, 3\}$ images by using Eq. (4) with a distinct focus depth d for each acquisition. Following Inagaki et al. [17], we set $d \in \{0\}$, $\{-1, 1\}$, and $\{-1, 0, 1\}$ for $N = 1$, 2, and 3, respectively. No trainable parameters were used for FS. In both CA and FS, the N acquired images were stacked along the channel dimension and fed directly to the reconstruction network. The sub-sampled image structure was not used because pixel-wise coding was not applied for CA or FS.

4.2.2 Snapshot Methods

We also compared TMCA and CFS with several snapshot methods, in which the entire LF $l(x, y, u, v)$ is acquired from a single observed image $i(x, y)$.

The first snapshot methods were two ablation cases of TMCA and CFS, in which $p_t(x, y)$ was simply disabled (i.e., $p_t(x, y)$ was set to 1); we denote these as **TMCA-** and **CFS-**, respectively.

We also considered an ideal case of freely designed coding, which is denoted as **Full 4D** and defined by

$$i(x, y) = \sum_{u,v} m(x, y, u, v) l(x, y, u, v), \tag{15}$$

where $m(u, v, x, y) \in [0, 1]$ can take any 4D modulation pattern that repeats every 8×8 pixels. Note that no hardware is available to implement this kind of free modulation; rather, this was only a software simulation. We implemented Eq. (15) with $5 \times 5 \times 8 \times 8$ trainable parameters. The acquired image $i(x, y)$ was also represented as a set of sub-sampled images with Eq. (14) and fed to the reconstruction network.

Next, we considered a hypothetical lens-array-based camera that was similar to a Lytro camera [4] (denoted as **Lytro-like**). This camera took 5×5 views with $W/5 \times H/5$ pixels. Let $l_{\downarrow}(x', y', u, v)$ be a spatially downscaled version of the original LF $l(x, y, u, v)$ with a ratio of $1/5 \times 1/5$. Then, the acquired image $i(x, y)$ was formulated as

$$i(5x' + u, 5y' + v) = l_{\downarrow}(x', y', u, v), \tag{16}$$

where $(x, y) = (5x' + u, 5y' + v)$. The angular coordinate (u, v) was interleaved in the cycle of 5×5 pixels on $i(x, y)$. Accordingly, we represented the observed image $i(x, y)$ as a set of 25 (5×5) sub-sampled images in a manner similar to that of Eq. (14):

$$i_{u,v}(x, y) = \begin{cases} i(x, y) & (x\%5, y\%5) = (u, v) \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

where each $i_{u,v}(x, y)$ contained only the information associated with a specific angular coordinate (u, v). All the sub-sampled images were stacked along the channel dimension and fed to the reconstruction network.

Finally, we mention an extreme case denoted as **Center-view**, in which only the central image was acquired and used for LF reconstruction. This approach is similar to some recent works on single-view view synthesis and monocular depth estimation [32], [33], in which the reconstruction network can use only a single image (without any coding) to obtain the 3D scene information. In our case, the acquired image was given by

$$i(x, y) = l(x, y, u_c, v_c), \tag{18}$$

which was fed directly to the reconstruction network.

Note that FS and CA could also be considered as snapshot methods when only a single image was used for reconstruction. In particular, the acquired image with FS ($N = 1$)

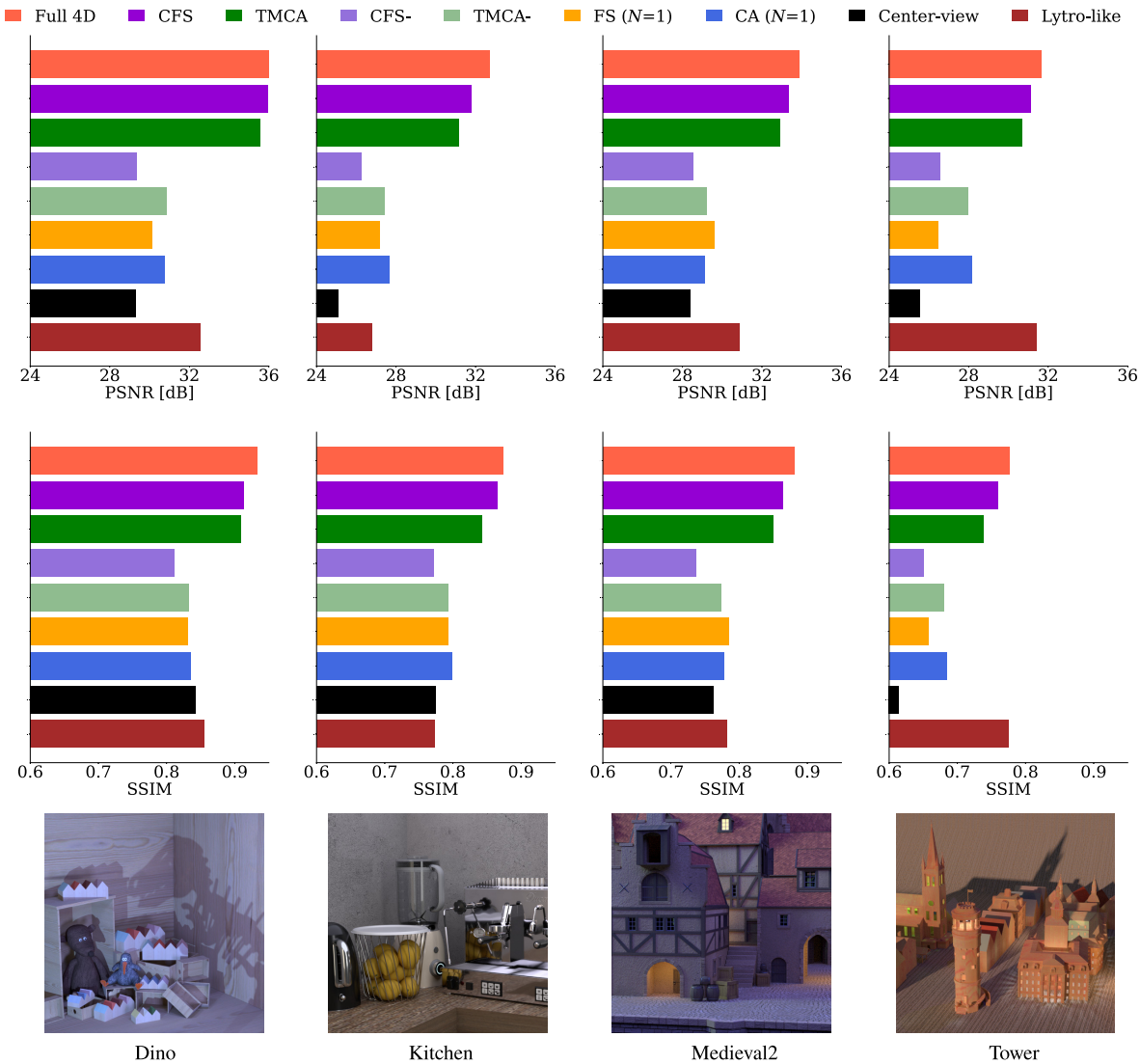


Fig. 6 Quantitative comparison among the snapshot methods for the Dino, Kitchen, Medieval2, and Tower LFs, from left to right. The PSNR (top) and SSIM (middle) scores are shown, along with the central view in each LF (bottom).

was given by Eq. (1), which served as the baseline representing the case without any coding.

4.2.3 Results

We first describe the comparisons with other possible methods for snapshot LF imaging, which included Full 4D, TMCA-, CFS-, Lytro-like, Center-view, CA ($N = 1$), and FS ($N = 1$). The quantitative results are summarized in Fig. 6. As expected, Full 4D achieved the best scores for all LFs. CFS and TMCA respectively yielded the second and third best scores on average. The difference between TMCA and CFS is discussed deeper in Sect. 4.4. The Lytro-like method obtained a good result for the Tower LF, possibly because this scene does not contain many high-frequency components, which are hard to obtain with the Lytro-like method. In contrast, both CFS and TMCA consistently achieved high quality for all four LFs.

In Fig. 7, we show the PSNR and SSIM scores for each viewpoint obtained with the Kitchen LF. Both TMCA and CFS yielded better quality in a stable manner across the viewpoints as compared to CA ($N = 1$) and FS ($N = 1$). Several visual results are presented in Fig. 8, in which we show the reconstructed top-left views with epipolar plane images (EPIs) and the difference from the ground truth (magnified by a factor of 3 for better visualization). From these results, we can see that TMCA and CFS respectively obtained visually more convincing results than those of the CA ($N = 1$) and FS ($N = 1$) methods from which they were derived.

We also evaluated the performance of TMCA and CFS with respect to CA and FS. Because the latter two methods can use multiple acquired images for LF reconstruction, the reconstruction quality is plotted against the number of ac-

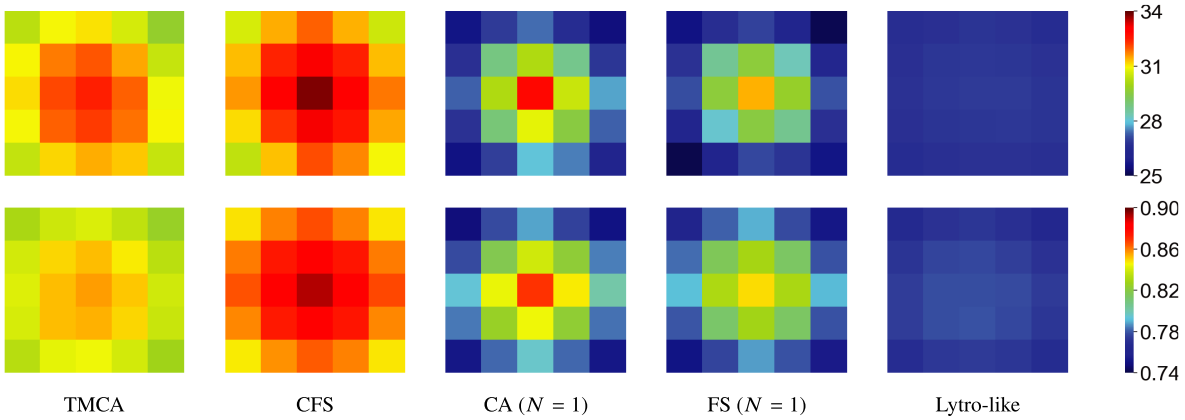


Fig. 7 PSNR (top) and SSIM (bottom) scores for each viewpoint with the Kitchen LF.

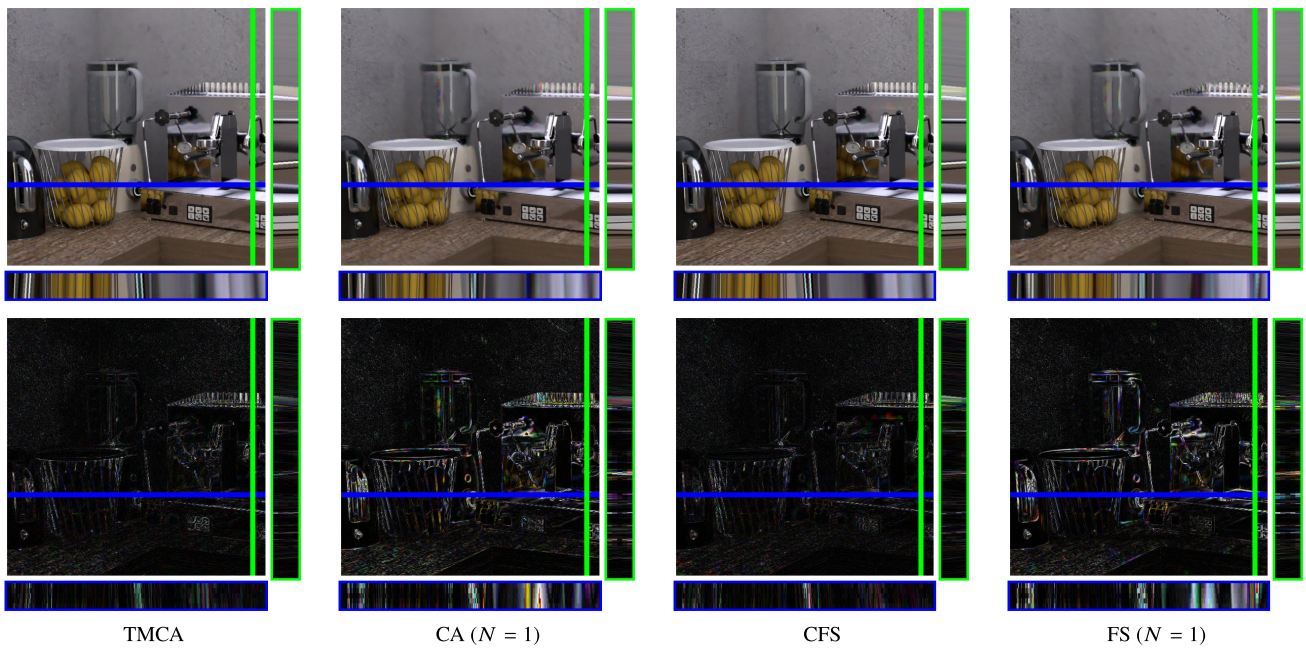


Fig. 8 Visual results for the Kitchen LF: reconstructed top-left views with EPIs (top), and differences from the ground truth (bottom, ×3 for better visualization).

quired images in Fig. 9. As mentioned above, TMCA and CFS clearly performed better than CA and FS as long as only a single acquired image was used. On the other hand, both CA and FS could significantly boost the reconstruction quality by increasing the number of acquired images. Notably, the scores obtained with TMCA and CFS were close to those obtained with CA and FS ($N = 2$). This result is understandable, because TMCA and CFS were designed to fuse T coded/focused images into a single observed image. Thus, it is natural that the scores for TMCA and CFS ($T = 2$ in this case) did not exceed but approached the scores for their multi-image counterparts (CA and FS with $N = 2$). Nevertheless, the results have shown that both TMCA and CFS performed reasonably well compared to the methods using multiple acquired images.

4.3 Evaluation of Network Architecture

We analyze the impact of the network architecture on the LF reconstruction task. We compared our network architecture with two other architectures that were developed for the CA method. Inagaki et al. [14] used a lightweight architecture that consisted of a stack of 2D convolutional layers and was similar to our architecture. Guo et al. [16] developed a more complicated network architecture that was designed specifically for the CA-based LF reconstruction task. Both of those architectures used the CA imaging method, in which the coding pattern for the aperture plane was jointly optimized with the reconstruction network. For this comparison, we retrained both networks by using the same dataset that we used for our network, with a configuration of $N = 1$

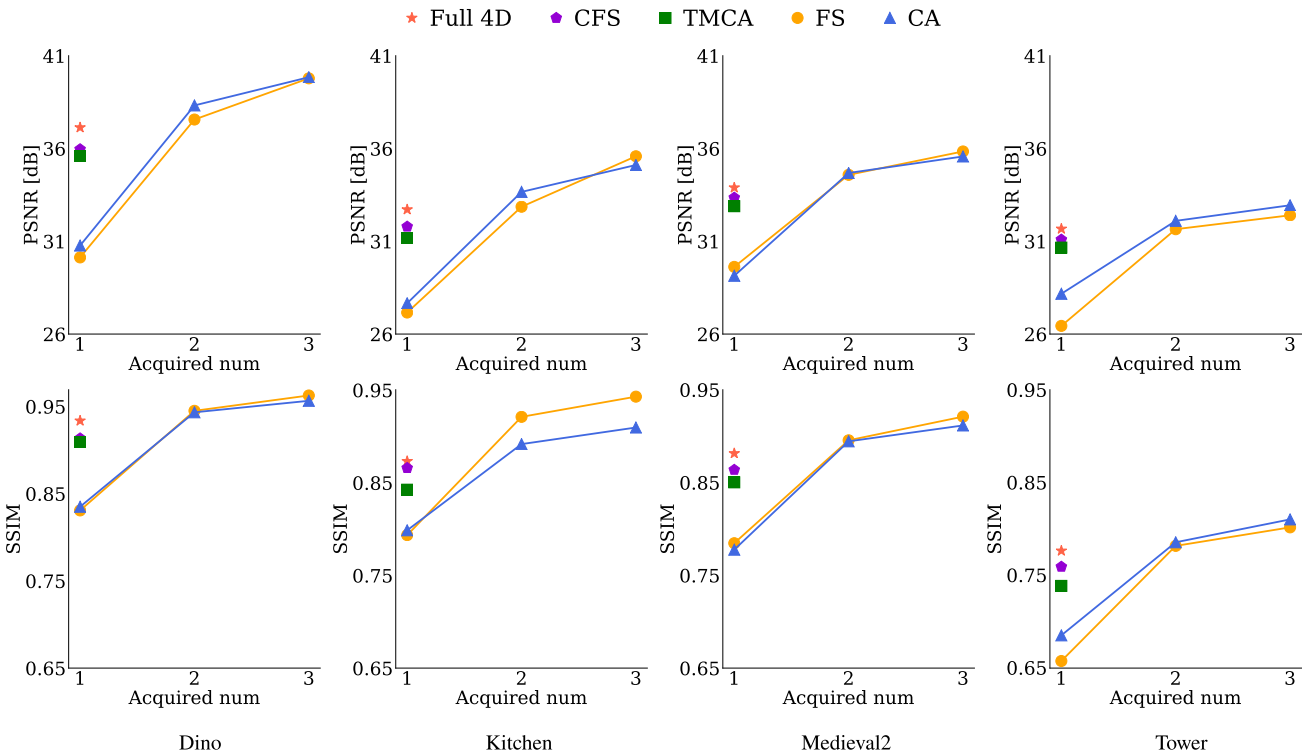


Fig. 9 Quantitative comparison of TMCA and CFS with respect to CA and FS. The PSNR (top) and SSIM (bottom) scores are shown for four LFs: Dino, Kitchen, Medieval2, and Tower, from left to right.

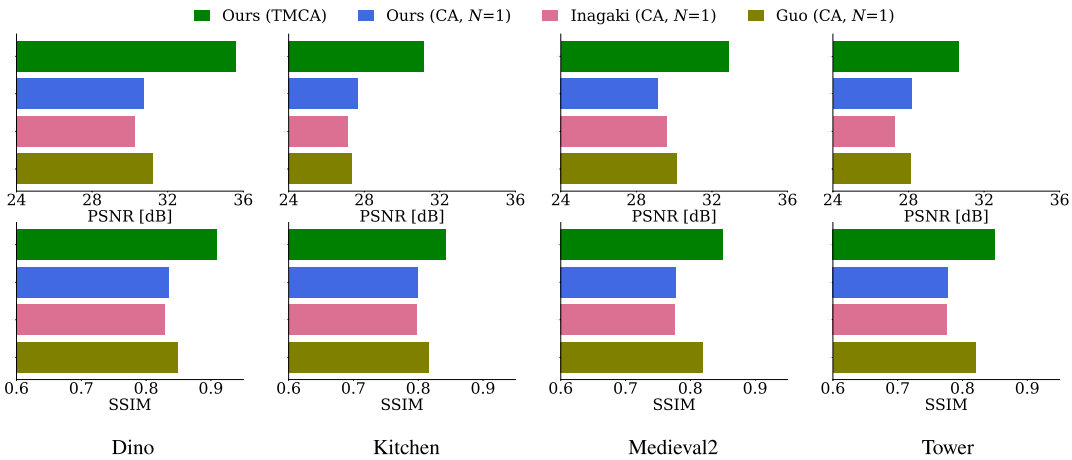


Fig. 10 Comparison of different reconstruction networks. PSNR (top) and SSIM (bottom) scores are presented for four LFs.

(i.e., a single image was used for LF reconstruction).

Figure 10 shows the reconstruction quality obtained by our network in comparison to Inagaki's and Guo's networks. Here, "ours (TMCA)" and "ours (CA, $N = 1$)" correspond to "TMCA" (with the sub-sampled image structure) and "CA with $N = 1$" (without the sub-sampled image structure) mentioned in Sect. 4.2. As expected, our network (CA, $N = 1$) and that of Inagaki (CA, $N = 1$) achieved almost equivalent quality on average. Meanwhile, Guo's network (CA, $N = 1$) yielded better reconstruction quality because of its sophisticated architecture. However, Guo's network did not reach the quality of ours (TMCA), which indicates the greater importance of the imaging method over the network architecture. Moreover, because of its complexity, Guo's network took 46.3 s to reconstruct a single LF, whereas ours (CA, $N = 1$), ours (TMCA), and Inagaki's network took only 145, 160, and 120 ms, respectively.

4.4 Detailed Analysis and Comparison of TMCA and CFS

As Fig. 6 shows, CFS achieved slightly better reconstruction quality than TMCA. Moreover, the quality difference

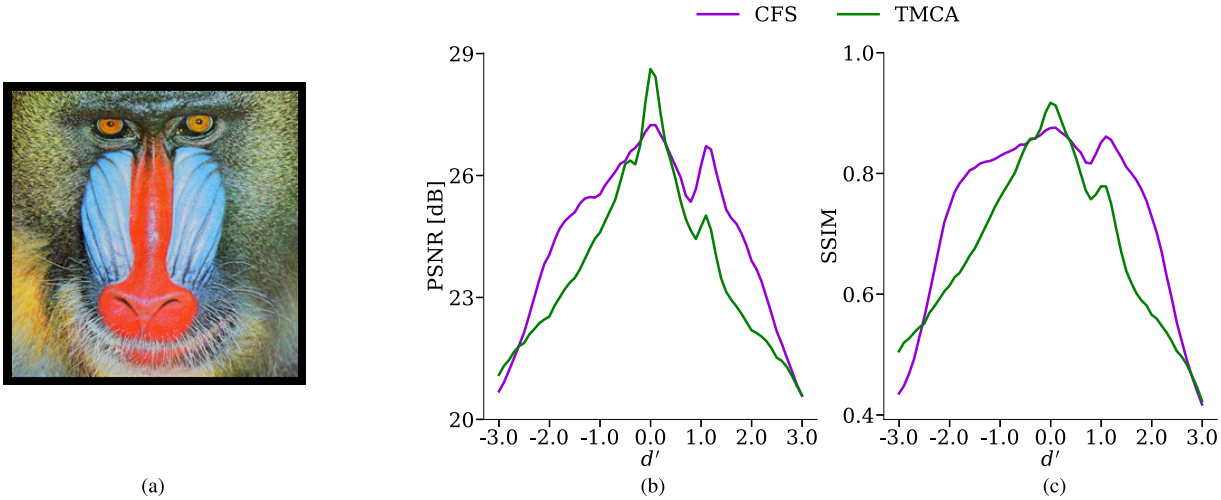


Fig. 11 Performance analysis of TMCA and CFS using planar scene; (a) texture used as $\mathcal{T}(x, y)$, and (b)(c) PSNR and SSIM scores plotted against disparity d'.

between CFS and TMCA seemingly depends on the target LFs. A possible factor behind these results is the depth (disparity) distribution of the target scene, as suggested by the analysis presented below.

We consider a hypothetical scene where a textured plane is located at a certain depth and it is facing straight to the camera. In this case, the entire scene has a constant disparity. The LF generated from the scene is described as

$$L_{d'}(x, y, u, v) = \mathcal{T}(x + d'(u - u_c), y + d'(v - v_c)) \quad (19)$$

where $\mathcal{T}(x, y)$ is the texture of the scene, d' [pixels/viewpoint] is the disparity, and (u_c, v_c) is the central viewpoint of the LF. Here, a larger disparity means a closer distance from the camera. We feed the generated LF $L_{d'}(x, y, u, v)$ to the imaging pipelines of TMCA and CFS that were pre-trained with the default parameters, and evaluate the resulting reconstruction quality of $L_{d'}(x, y, u, v)$. We conduct this evaluation with different values for d' to find depth-dependent characteristics of TMCA and CFS.

We used an image with 272 × 272 pixels shown in Fig. 11 (a) as the texture $\mathcal{T}(x, y)$. We varied d' in the range of $[-3, 3]$, because with the LFs used in Fig. 6, the disparity values mostly fall within this range. The PSNR and SSIM scores for the reconstructed $L_{d'}(x, y, u, v)$ are plotted against d' in Fig. 11(b)(c).

As an overall trend for both TMCA and CFS, the reconstruction quality had its peak around $d' = 0$, and tended to decline as d' diverged from 0. We also observe that CFS outperformed TMCA for most of the values of d', whereas TMCA was better than CFS only in the narrow ranges around $d' = 0$ and $d' = -3$. We guess that the better performance of CFS over various disparity values can be attributed to the shear operation, where the target LF is "aligned" or "focused" for various disparity values (see Eq. (4)). Meanwhile, TMCA yielded a sharper performance peak around $d' = 0$, because it involved no shear operation.

We can draw several insights from this analysis, considering that the disparity is position variant (i.e., d' can take different values pixel to pixel) for a general scene. First, it is more likely that CFS performs better than TMCA for a general scene, because the disparity can take various values not limited to the vicinities of $d' = 0$ and $d' = -3$. Moreover, the quality difference between CFS and TMCA will depend on the depth distribution of the target scene; for example, if the target scene has more objects around $d' = 0$, it would be more favorable to TMCA. Finally, as the target scene contains more objects with larger $|d'|$ values, its accurate reconstruction would become more difficult for both TMCA and CFS.

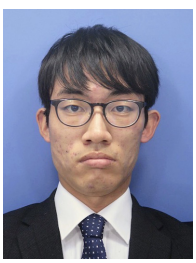
5. Conclusion

In this paper, we considered snapshot compressive LF imaging, in which an entire LF is obtained from only one acquired image. We focused on two promising imaging methods, TMCA and CFS, both of which involve rapid optical control during an exposure. We developed a unified algorithm pipeline to enable implementation and comparison of several methods for compressive LF imaging under the same conditions. We found that both TMCA and CFS achieved better reconstruction quality than other possible snapshot methods, and they also performed reasonably well in comparison to methods using multiple acquired images. We believe that our work will contribute to extension of the framework for compressive LF imaging to cover various imaging methods based on different camera architectures. Our future work will include exploration of better network architectures and evaluation of TMCA and CFS on real imaging hardware. Extension to moving scenes (LFs moving over time) [27] will also be an interesting avenue for future work.

References

[1] E.H. Adelson and J.R. Bergen, "The plenoptic function and the elements of early vision," Computational Models of Visual Processing,

pp.3–20, 1991.
[2] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen, "The lumigraph," Proc. 23th Annual Conf. Computer Graphics and Interactive Techniques, pp.43–54, Aug. 1996.
[3] A. Isaksen, L. McMillan, and S.J. Gortler, "Dynamically reparameterized light fields," Proc. 27th Annual Conf. Computer Graphics and Interactive Techniques, pp.297–306, July 2000.
[4] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Computer Science Technical Report, vol.2, no.11, pp.1–11, 2005.
[5] B. Mildenhall, P.P. Srinivasan, R. Ortiz-Cayon, N.K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," ACM TOG, vol.38, no.4, pp.1–14, Aug. 2019.
[6] T.C. Wang, A.A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," IEEE Trans. PAMI, vol.38, no.11, pp.2170–2181, Nov. 2016.
[7] C. Shin, H. Jeon, Y. Yoon, I. Kweon, and S. Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp.4748–4757, 2018.
[8] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting," ACM TOG, vol.31, no.4, pp.1–11, July 2012.
[9] S. Lee, C. Jang, S. Moon, J. Cho, and B. Lee, "Additive light field displays: realization of augmented reality with holographic optical elements," ACM TOG, vol.35, no.4, Article No. 60, July 2016.
[10] B. Wilburn, N. Joshi, V. Vaish, E.V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," ACM TOG, vol.24, no.3, pp.765–776, July 2005.
[11] T. Fujii, K. Mori, K. Takeda, K. Mase, M. Tanimoto, and Y. Suenaga, "Multipoint measuring system for video and sound-100-camera and microphone system," IEEE Int. Conf. Multimedia and Expo (ICME), pp.437–440, 2006.
[12] M. Levoy and P. Hanrahan, "Light field rendering," Proc. 23th Annual Conf. Computer Graphics and Interactive Techniques, pp.31–42, Aug. 1996.
[13] H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S.K. Nayar, "Programmable aperture camera using LCoS," European Conf. Comput. Vis. (ECCV), pp.337–350, 2010.
[14] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara, "Learning to capture light fields through a coded aperture camera," European Conf. Comput. Vis. (ECCV), pp.431–448, 2018.
[15] A.K. Vadathya, S. Girish, and K. Mitra, "A unified-learning based framework for light field reconstruction from coded projections," IEEE Trans. Comput. Imag., pp.304–316, 2019.
[16] M. Guo, J. Hou, J. Jin, J. Chen, and L.P. Chau, "Deep spatial-angular regularization for compressive light field reconstruction over coded apertures," European Conf. Comput. Vis. (ECCV), pp.278–294, 2020.
[17] Y. Inagaki, K. Takahashi, and T. Fujii, "Light field acquisition from focal stack via a deep CNN," International Display Workshop (IDW), pp.1077–1080, 2019.
[18] K. Takahashi, Y. Kobayashi, and T. Fujii, "From focal stack to tensor light-field display," IEEE Trans. Image Process., vol.27, no.9, pp.4571–4584, Sept. 2018.
[19] K. Tateishi, K. Sakai, C. Tsutake, K. Takahashi, and T. Fujii, "Factorized modulation for singleshot lightfield acquisition," IEEE Int. Conf. Image Process. (ICIP), pp.3253–3257, 2021.
[20] E. Vargas, J.N. Martel, G. Wetzstein, and H. Arguello, "Time-multiplexed coded aperture imaging: Learned coded aperture and pixel exposures for compressive imaging systems," IEEE Int. Conf. Comput. Vis. (ICCV), pp.2692–2702, 2021.
[21] X. Lin, J. Suo, G. Wetzstein, Q. Dai, and R. Raskar, "Coded focal stack photography," IEEE Int. Conf. Computational Photography (ICCP), pp.1–9, 2013.
[22] D. Liu, J. Gu, Y. Hitomi, M. Gupta, T. Mitsunaga, and S.K. Nayar, "Efficient space-time sampling with pixel-wise coded exposure for high-speed imaging," IEEE Trans. PAMI, vol.36, no.2, pp.248–260, 2013.
[23] M. Yoshida, T. Sonoda, H. Nagahara, K. Endo, Y. Sugiyama, and R.I. Taniguchi, "High-speed imaging using CMOS image sensor with quasi pixel-wise exposure," IEEE Trans. Comput. Imag., vol.6, pp.463–476, 2019.
[24] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp.1831–1838, 2010.
[25] K. Kodama and A. Kubota, "Efficient reconstruction of all-in-focus images through shifted pinholes from multi-focus images for dense light field synthesis and rendering," IEEE Trans. Image Process., vol.22, no.11, pp.4407–4421, Nov. 2013.
[26] H. Nagahara, S. Kuthirummal, C. Zhou, and S.K. Nayar, "Flexible depth of field photography," European Conf. Comput. Vis. (ECCV), pp.60–73, 2008.
[27] K. Sakai, K. Takahashi, T. Fujii, and H. Nagahara, "Acquiring dynamic light fields through coded aperture camera," European Conf. Comput. Vis. (ECCV), pp.368–385, 2020.
[28] Computer Graphics Laboratory, Stanford University, "The (new) stanford light field archive," 2018. http://lightfield.stanford.edu.
[29] MIT Media Lab's Camera Culture Group, "Compressive light field camera," 2015. http://cameraculture.media.mit.edu/projects/compressive-light-field-camera/.
[30] Heidelberg Collaboratory for Image Processing, "Datasets and benchmarks for densely sampled 4D light fields," 2016. http://lightfieldgroup.iwr.uni-heidelberg.de/?page_id=713.
[31] Heidelberg Collaboratory for Image Processing, "4D light field dataset," 2018. http://hci-lightfield.iwr.uni-heidelberg.de/.
[32] P.P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4D RGBD light field from a single image," European Conf. Comput. Vis. (ECCV), pp.2262–2270, 2017.
[33] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, "Unsupervised monocular depth estimation from light field image," IEEE Trans. Image Process., vol.29, pp.1606–1617, 2019.



Kohei Tateishi received B.E. and M.E. degrees in information and communication engineering from Nagoya University, Japan, in 2020 and 2022. When he was a student, his research topic was compressive light-field imaging.

Chihiro Tsutake received B.E., M.E., and Ph.D. degrees from the University of Fukui, Japan, in 2015, 2017, and 2020. Since 2020, he has been with the Graduate School of Engineering, Nagoya University, as an assistant professor. His research interests include image/video compression, image restoration, and 3D image processing.

Keita Takahashi received B.E., M.S., and Ph.D. degrees in information and communication engineering from the University of Tokyo, Japan, in 2001, 2003, and 2006. He was a project assistant professor at the University of Tokyo from 2006 to 2011 and an assistant professor at the University of Electro-Communications from 2011 to 2013. Since 2013, he has been with the Graduate School of Engineering, Nagoya University, as an associate professor. His research interests include image processing, computational photography, and 3D displays. He is a member of the IEEE Computer Society and Signal Processing Society, the Information Processing Society of Japan, and the Institute of Image Information and Television Engineers of Japan.

Toshiaki Fujii received B.E., M.E., and Dr.E. degrees in electrical engineering from the University of Tokyo, Japan, in 1990, 1992, and 1995. In 1995, he joined the Graduate School of Engineering, Nagoya University, where he is currently a professor. From 2008 to 2010, he was with the Graduate School of Science and Engineering, Tokyo Institute of Technology. From 2019 to 2021, he also served as a senior science and technology policy fellow of the Cabinet Office, Government of Japan. His current research interests include multidimensional signal processing, multi-camera systems, multi-view video coding and transmission, free-viewpoint video, and their applications. He is a member of the IEEE Signal Processing Society, the ISO/IEC JTC1/SC29/WG4, WG1 (MPEG-I Visual, JPEG) standardization committee of Japan, the Information Processing Society of Japan, and the Institute of Image Information and Television Engineers of Japan.