Fast Light Field Reconstruction with Deep Coarse-to-Fine Modeling of Spatial-Angular Clues

Henry Wing Fung Yeung[1], Junhui Hou[2(✉)], Jie Chen[3], Yuk Ying Chung[1], and Xiaoming Chen[4]

[1] School of Information Technologies, University of Sydney, Sydney, Australia
[2] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
jh.hou@cityu.edu.hk
[3] School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore, Singapore
[4] School of Information Science and Technology, University of Science and Technology of China, Hefei, China

Abstract. Densely-sampled light fields (LFs) are beneficial to many applications such as depth inference and post-capture refocusing. However, it is costly and challenging to capture them. In this paper, we propose a learning based algorithm to reconstruct a densely-sampled LF fast and accurately from a sparsely-sampled LF in one forward pass. Our method uses computationally efficient convolutions to deeply characterize the high dimensional spatial-angular clues in a coarse-to-fine manner. Specifically, our end-to-end model first synthesizes a set of intermediate novel sub-aperture images (SAIs) by exploring the coarse characteristics of the sparsely-sampled LF input with spatial-angular alternating convolutions. Then, the synthesized intermediate novel SAIs are efficiently refined by further recovering the fine relations from all SAIs via guided residual learning and stride-2 4-D convolutions. Experimental results on extensive real-world and synthetic LF images show that our model can provide more than 3 dB advantage in reconstruction quality in average than the state-of-the-art methods while being computationally faster by a factor of 30. Besides, more accurate depth can be inferred from the reconstructed densely-sampled LFs by our method.

Keywords: Light field · Deep learning
Convolutional neural network · Super resolution · View synthesis

1 Introduction

Compared with traditional 2-D images, which integrate the intensity of the light rays from all directions at a pixel location, LF images separately record the light

H.W.F. Yeung and J. Hou—Equal Contributions.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11210, pp. 138–154, 2018.
https://doi.org/10.1007/978-3-030-01231-1_9

ray intensity from different directions, thus providing additional information on the 3-D scene geometry. Such information is proportional to the angular resolution, i.e. the number of directions of the light rays, captured by the LF image. Densely sampled LF, with high resolution in the angular domain, contains sufficient information for accurate depth inference [1–4], post-capture refocusing [5] and 3D display [6,7].

LF images [8,9] can be acquired in a single shot using camera arrays [10] and consumer hand-held LF cameras such as Lytro [11] and Raytrix [12]. The former, due to the large number of sensors, can capture LF with higher spatial resolution while being expensive and bulky. Through multiplexing the angular domain into the spatial domain, the later is able to capture LF images with a single sensor, and thus are cheaper and portable. However, due to the limited sensor resolution, there is a trade-off between spatial and angular resolution. As a result, these cameras cannot densely sample in both the spatial and angular domains.

Reconstruction of a densely-sampled LF from a sparsely-sampled LF input is an on-going problem. Recent development in deep learning based LF reconstruction models [13,14] have achieved far superior performance over the traditional approaches [1–4]. Most notably, Kalantari *et al.* [13] proposed a sequential convolutional neural network (CNN) with disparity estimation and Wu *et al.* [14] proposed to use a blur-deblur scheme to counter the problem of information asymmetry between angular and spatial domain and a single CNN is used to map the blurred epipolar-plane images (EPIs) from low to high resolution. However, both approaches require heavy pre- or post-processing steps and long runtime, making them impractical to be applied in consumer LF imaging system.

In this paper, we propose a novel learning based model for fast reconstruction of a densely-sampled LF from a very sparsely-sampled LF. Our model, an end-to-end CNN, is composed of two phases, i.e., view synthesis and refinement, which are realized by computationally efficient convolutions to deeply characterize the spatial-angular clues in a coarse-to-fine manner. Specifically, the view synthesis network is designed to synthesize a set of intermediate novel sub-aperture images (SAIs) based on the input sparsely-sampled LF and the view refinement network is deployed for further exploiting the intrinsic LF structure among the synthesized novel SAIs. Our model does not require disparity warping nor any computationally intensive pre- and post-processing steps. Moreover, reconstruction of all novel SAIs are performed in one forward pass during which the intrinsic LF structural information among them is fully explored. Hence, our model fully preserves the intrinsic structure of reconstructed densely-sampled LF, leading to better EPI quality that can contribute to more accurate depth estimation.

Experimental results show that our model provides over 3 dB improvement in the average reconstruction quality while requiring less than 20 s on CPU, achieving over 30× speed up, compared with the state-of-the-art methods in synthesizing a densely-sampled LF from a sparsely-sampled LF. Experiment also shows that the proposed model can perform well on large baseline LF inputs and provides substantial quality improvement of over 3 dB with extrapolation. Our

algorithm not only increases the number of samples for depth inference and post-capture refocusing, it can also enable LF to be captured with higher spatial resolution from hand-held LF cameras and potentially be applied in compression of LF images.

2 Related Work

Early works on LF reconstruction are based on the idea of warping the given SAIs to novel SAIs guided by an estimated disparity map. Wanner and Goldluecke [1] formulated the SAI synthesis problem as an energy minimization problem with a total variation prior, where the disparity map is obtained through global optimisation with a structure tensor computed on the 2-D EPI slices. Their approach considers disparity estimation as a separate step from SAI synthesis, which makes the reconstructed LF heavily dependent on the quality of the estimated disparity maps. Although subsequent research [2–4] has shown significantly better disparity estimations, ghosting and tearing effects are still present when the input SAIs are sparse.

Kalantari *et al.* [13] alleviated the drawback of Wanner and Goldluecke [1] by synthesizing the novel SAIs with two sequential CNNs that are jointly trained end-to-end. The first CNN performs disparity estimation based on a set of depth features pre-computed from the given input SAIs. The estimated disparities are then used to warp the given SAIs to the novel SAIs for the second CNN to perform color estimation. This approach is accurate but slow due to the computation intensive depth features extraction. Furthermore, each novel SAI is estimated at a separate forward pass, hence the intrinsic LF structure among the novel SAIs is neglected. Moreover, the reconstruction quality depends heavily upon the intermediate disparity warping step, and thus the synthesized SAIs are prone to occlusions.

Advancement in single image super-resolution (SISR) is recently made possible by the adoption of deep CNN models [15–18]. Following this, Yoon *et al.* [19,20], developed a CNN model that jointly super-resolves the LF in both the spatial and angular domain. This model concatenates at the channel dimension a subset of the spatially super-resolved SAIs from a CNN that closely resembles the model proposed in [15]. The concatenated SAIs are then passed into a second CNN for angular super-resolution. Their approach is designed specificity for scale 2 angular super-resolution and can not flexibly adapt to perform on very sparsely-sampled LF input.

Recently, Wu *et al.* [14] developed a CNN model that inherits the basic architecture of [15] with an addition residual learning component as in [16]. Using the idea of SISR, their model focuses on recovering the high frequency details of the bicubic upsampled EPI while a blur-deblur scheme is proposed to counter the information asymmetry problem caused by sparse angular sampling. Their model is adaptable to different devices. Since each EPI is a 2-D slice in both the spatial and angular domains of the 4-D LF, EPI based model can only utilize

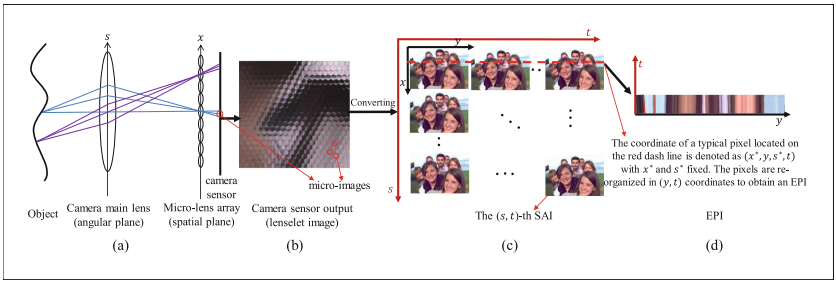


Fig. 1. LF captured with a single sensor device. The angular information of an LF is captured via the separation of light rays by the micro-lens array. The resulting LF can be parameterized by the spatial coordinates and the angular coordinates, i.e. the position of the SAI.

SAIs from the same horizontal or vertical angular coordinate of the sparsely-sampled LF to recover the novel SAIs in between, thus severely restricting the accessible information of the model. For the novel SAIs that do not fall within the same horizontal or vertical angular coordinate as the input SAIs, they are reconstructed based on the previously estimated SAIs. As a result, these SAIs are biased due to input errors. Moreover, due to the limitation in the blurring kernel size and bicubic interpolation, this method cannot be applied to sparsely-sampled LF with only 2×2 SAIs or with disparity larger than 5 pixels.

3 The Proposed Approach

3.1 4-D Light Field and Problem Formulation

4-D LF can be represented using the two-plane parameterization structure, as illustrated in Fig. 1, where the light ray travels and intersects the angular plane (s, t) and the spatial plane (x, y) [21]. Let $\mathcal{I} \in \mathbb{R}^{W \times H \times M \times N \times 3}$ denote an LF with $M \times N$ SAIs of spatial dimension $W \times H \times 3$, and $\mathcal{I}(:, :, s, t, :) \in \mathbb{R}^{W \times H \times 3}$ be the (s, t)-th SAI $(1 \leq s \leq M,\ 1 \leq t \leq N)$.

Densely-sampled LF reconstruction aims to construct an LF $\mathcal{I}' \in \mathbb{R}^{W \times H \times M' \times N' \times 3}$ including a large number of SAIs, from an LF $\mathcal{I}$ containing a small number of SAIs , where $M' > M$ and $N' > N$. Since the densely-sampled LF $\mathcal{I}'$ also contains the set of input SAIs, denoted as $\mathcal{K}$, the SAIs to be estimated is therefore reduced to the set of $(M' \times N' - M \times N)$ novel SAIs, denoted as $\mathcal{N}$.

Efficient modelling of the intrinsic structure of LF , i.e. photo-consistency, defined as the relationship of pixels from different SAIs that represent the same scene point, is crucial for synthesising high quality LF SAIs. However, real-world scenes usually contain factors such as occlusions, specularities and

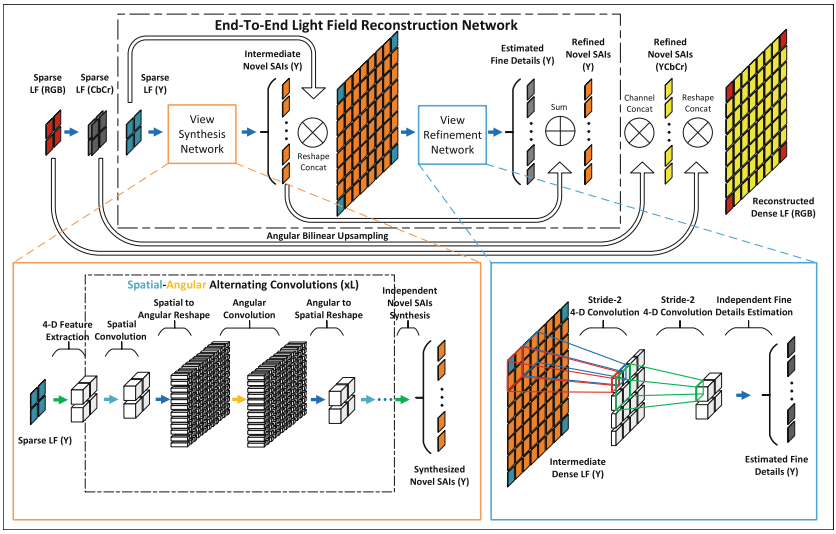


Fig. 2. The workflow of reconstructing a densely-sampled LF with 8×8 SAIs from a sparsely-sampled LF with 2×2 SAIs. Our proposed model focuses on reconstructing the luma components (Y) of the novel SAIs, while angular bilinear interpolation recovers the other two chrominance components (Cb and Cr). Note that the reshape operations in the view synthesis network are included for understanding of the data flow and are not required in actual implementation.

non-Lambertian lighting, making it challenging to characterize this structure accurately. In this paper, we propose a CNN based approach for efficient characterisation of spatial-angular clues for high quality reconstruction of densely sampled LFs.

3.2 Overview of Network Architecture

As illustrated in Fig. 2, we propose a novel CNN model to provide direct end-to-end mapping between the luma component of the input SAIs, denoted as $\mathcal{K}_Y$, and that of the novel SAIs, denoted as $\widehat{\mathcal{N}}_Y$. Our proposed network consists of two phases: view synthesis and view refinement. The view synthesis network, denoted as $f_S(.)$, first synthesizes the whole set of intermediate novel SAIs based on all input SAIs. The synthesized novel SAIs are then combined with the input SAIs to form a 4-D LF structure using a customised reshape-concat layer. This intermediate LF is then fed into the refinement network, denoted as $f_R(.)$, for recovering the fine details. At the end, the estimated fine details are added to the intermediate synthesized SAIs in an pixel-wise manner to give the final prediction of the novel SAIs $\widehat{\mathcal{N}}_Y$. The relations between the inputs and outputs of our model is represented as:

$$\widehat{\mathcal{N}}_Y = f_S(\mathcal{K}_Y) + f_R(f_S(\mathcal{K}_Y), \mathcal{K}_Y). \tag{1}$$

Note that the full color novel SAIs $\widehat{\mathcal{N}}$ are obtained from combining $\widehat{\mathcal{N}}_Y$ with angular bilinear interpolation of the other two chrominance components, i.e., Cb and Cr. Contrary to the previous approaches that synthesize a particular novel SAI at a each forward pass [13], and an EPI of a row or column of novel SAIs at each forward pass [14], our approach is capable of jointly producing all novel SAIs at one pass to preserve the intrinsic LF structure among them. Our network is full 4-D convolutional and uses Leaky Relu with the parameter of 0.2 for activation. Table 1 provides a summary of the network architecture.

3.3 View Synthesis Network

The view synthesis network estimates a set of intermediate novel SAIs by uncovering the coarse spatial-angular clues carried by the limited number of SAIs of the input sparsely-sampled LF. This step takes in all input SAIs from the given LF for the estimation of novel SAIs, and thus it can make full use of available information on the structural relationship among SAIs. For achieving this, it is necessary to perform convolution on all both the spatial and the angular dimensions of the input LF.

4-D convolution is a straightforward choice for this task. However, for this particular problem, the computational cost required by 4-D convolution makes training such a model impossible in a reasonable amount of time. Pseudo filters or separable filters, which reduce model complexity by approximating a high dimensional filter with filters of lower dimension, have been applied to solve different computer vision problems, such as image structure extraction [22], 3-D rendering [23] and video frame interpolation [24]. This is recently adopted in [25] for LF material classification, which verifies that the pseudo 4-D filters can achieve similar performance as 4-D filters.

For preventing potential overfitting and long training time from the use of full 4-D filter while characterizing 4-D information of LF, we adopt the pseudo 4-D filter which approximates a single 4-D filtering step with two 2-D filters that perform convolution on the spatial and the angular dimensions of the LF in an alternating manner. Such a design requires only the computation of $2/n^2$ of a 4-D convolution while still utilizing all available information from the input SAIs.

In the synthesis network, spatial-angular alternating convolutions are adopted only for intermediate feature extraction. For the initial feature extraction step and the novel SAIs synthesis step, 4-D convolution is applied since the computational complexity is less. Such a design obtains a significant reduction in parameter size as well as computational cost. Moreover, the low computational cost also benefits from that feature extraction is performed at the coarse angular resolution of $M \times N$ as opposed to [14] at the fine level of $M' \times N'$.

Table 1. Model specification for reconstructing a densely-sampled LF with 8×8 SAIs from a sparsely-sampled LF with 2×2 SAIs on the luma component. The first two dimensions of the filters, input and output data tensor correspond to the spatial dimension whereas the third and the forth dimension correspond to the angular dimension. The fifth dimension of the output tensor denotes the number of feature maps in the intermediate convolutional layers while representing the number of novel SAIs at the final layer. Stride and Paddings are given in the form of (Spatial/Angular). All convolutional layers contain biases. Note that the intermediate LF reconstruction step is performed with reshape and concatenation operations to enable back-propagation of loss from the view refinement network to the view synthesis network.

	Filter size/operation	Input Size	Output Size	Stride	Pad
sparsely-sampled LF input	-	-	$(64, 64, 2, 2, 1)$	-	-
View synthesis netowrk					
Feature extraction	$(3, 3, 3, 3, 1, 64)$	$(64, 64, 2, 2, 1)$	$(64, 64, 2, 2, 64)$	1/1	1/1
Alternating filtering $(\times L)$					
Spatial $S_l, l \in \{1, ..., L\}$	$(3, 3, 1, 1, 64, 64)$	$(64, 64, 2, 2, 64)$	$(64, 64, 2, 2, 64)$	1/1	1/0
Angular $A_l, l \in \{1, ..., L\}$	$(1, 1, 3, 3, 64, 64)$	$(64, 64, 2, 2, 64)$	$(64, 64, 2, 2, 64)$	1/1	0/1
Novel SAIs synthesis	$(3, 3, 2, 2, 64, 60)$	$(64, 64, 2, 2, 64)$	$(64, 64, 1, 1, 60)$	1/1	1/0
Intermediate LF Reconstruction	Reshape & concat	$(64, 64, 2, 2, 1)$ $(64, 64, 1, 1, 60)$	$(64, 64, 8, 8, 1)$	-	-
View refinement network					
Angular Dim. Reduction 1	$(3, 3, 2, 2, 1, 16)$	$(64, 64, 8, 8, 1)$	$(64, 64, 4, 4, 16)$	1/2	1/0
Angular Dim. Reduction 2	$(3, 3, 2, 2, 16, 64)$	$(64, 64, 4, 4, 16)$	$(64, 64, 2, 2, 64)$	1/2	1/0
Fine details recovery	$(3, 3, 2, 2, 64, 60)$	$(64, 64, 2, 2, 64)$	$(64, 64, 1, 1, 60)$	1/1	1/0
Novel SAIs reconstruction	Element-wise sum	$(64, 64, 1, 1, 60)$ $(64, 64, 1, 1, 60)$	$(64, 64, 1, 1, 60)$	-	-

3.4 View Refinement Network

In the view synthesis phase, novel SAIs are independently synthesized, and the relationship among them is not taken into account. Therefore, a view refinement network is designed to further exploit the relationship among the synthesized novel SAIs from the intermediate LF, which is expected to contribute positively to the reconstruction quality of the densely-sampled LF. This can be considered as a regularizer that imposes the LF structure on the synthesized SAIs.

Inspired by the success of residual learning on image reconstruction [14,16–18], we equip our view refinement network with guided residual learning that is specifically designed for the LF data structure. Typical residual learning attempts to learn a transformation $R(\cdot)$ to recover the residual $R(\mathcal{I}')$ for the input data $\mathcal{I}'$, i.e. the intermediate LF, as shown in Eq. (2). However, the input to the refinement network consists of a set of SAIs $\mathcal{K}_Y \subset \mathcal{I}'$ from the given sparsely-sampled LF, which is absolutely correct, i.e. $R(\mathcal{K}_Y) = 0$, and a set of synthesized SAIs $\mathcal{N}'_Y = f_S(\mathcal{K}_Y) \subset \mathcal{I}'$, which is erroneous. Hence, residual learning on $\mathcal{K}_Y$ is unnecessary. Guided residual learning can be formulated as a typical residual learning on $\mathcal{N}'_Y$ with the guidance from the additional input, $\mathcal{K}_Y$, as shown in Eq. (3).

$$\widehat{\mathcal{I}}_Y = \mathcal{I}' + R(\mathcal{I}') \tag{2}$$

$$\widehat{\mathcal{N}}_Y = \mathcal{N}'_Y + R(\mathcal{N}'_Y|\mathcal{K}_Y) \tag{3}$$

Guided residual learning has the following benefits: (i) $\mathcal{K}_Y$, as a set of ground-truth SAIs, offers correct complementary information of the scene; (ii) learning 0 residuals for $\mathcal{K}_Y$ is not performed; and (iii) By placing $\mathcal{K}_Y$ and $\mathcal{N}'_Y$ in the form of $\mathcal{I}'$, a densely sampled intermediate LF, for input to the second stage refinement network, it encourages the first stage, i.e., view synthesis network, to generate SAIs that preserve the LF structure exhibiting in the EPI shown in Fig. 1(d).

Since the angular dimension increases significantly from $M \times N$ to $M' \times N'$ after the view synthesis processes, alternating convolution will incur a substantially higher computation cost that increases linearly in angular dimension. For reducing the computation to a manageable level, stride-2 4-D convolution is used for efficient angular dimension reduction while the feature map number is set to increase gradually. Note that to allow back-propagation, an intermediate 4-D LF is reconstructed from the previously synthesized novel SAIs and the input SAIs via a customised reshape-concat layer. The refinement details of all novel SAIs are independently estimated at the final 4-D convolution layer and are added to the previously synthesized intermediate novel SAIs to give the final reconstructed novel SAIs.

3.5 Training Details

The training objective is to minimise the $\mathcal{L}_2$ distance between all reconstructed novel SAIs $\widehat{\mathcal{N}}_Y$ and their respective ground-truth $\mathcal{N}_Y$:

$$\mathcal{L}_2(\mathcal{N}_Y, \widehat{\mathcal{N}}_Y) = \sum_x \sum_y \sum_s \sum_t \left(\widehat{\mathcal{N}}_Y(x,y,s,t) - \mathcal{N}_Y(x,y,s,t)\right)^2.$$

We trained a model for each task on the training set with 100 scenes provided by Kalantari *et al.* [13][1]. All images were taken with a Lytro Illum camera and were decoded to 14×14 SAIs with spatial resolution 376×541. Since the three SAIs from each side are usually black, we only adopted the middle 8×8 SAIs for training and testing as done in [13].

Training LFs were spatially cropped to 64×64 patches with stride 1, giving a maximum of approximately 15,000,000 training samples. Moreover, we adopted stochastic gradient descent to optimize the model, and the batch size was set to 1. The spatial resolution of the model output is kept unchanged at 64×64 with padding of zeros.

We implemented the model with the MatConvNet toolbox [26] in MATLAB and trained it with the GTX 1080 Ti GPU. Random filter weights under the MSRA method [27] were used to initialize our model, while biases were initialized to 0. Throughout training, momentum parameter was set to 0.9. Depending

[1] http://cseweb.ucsd.edu/~viscomp/projects/LF/papers/SIGASIA16.

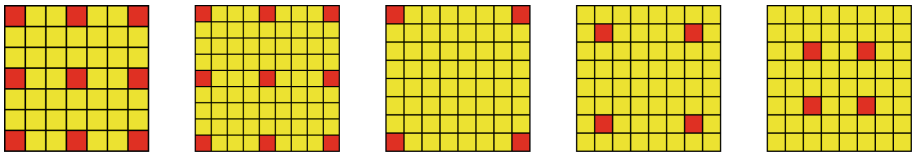


Fig. 3. Illustration of inputs (red blocks) and outputs (yellow blocks) for different tasks. From left to right: (a) $3 \times 3 - 7 \times 7$, (b) $3 \times 3 - 9 \times 9$, (c) $2 \times 2 - 8 \times 8$, (d) $2 \times 2 - 8 \times 8$ extrapolation-1, (e) $2 \times 2 - 8 \times 8$ extrapolation-2. (Color figure online)

on model depth, a learning rate between $1e{-}6$ to $2e{-}5$ was applied without weight decay, and epoch number was set between 8000 to 12000 each with 1000 iterations. Training time increases linearly with the number of alternating convolutions, ranging from around 1 day for model with 1 alternating convolution and 10 days for model with 16 alternating convolutions.

4 Experimental Results

Our model was compared with two state-of-the-art CNN based methods that are specifically designed for densely-sampled LF reconstruction, i.e., Kalantari *et al.* [13] and Wu *et al.* [14]. Comparisons were performed over three different tasks, shown in Fig. 3: $3 \times 3 - 7 \times 7$, $3 \times 3 - 9 \times 9$ and $2 \times 2 - 8 \times 8$. Task $M \times N - M' \times N'$ stands for reconstructing densely-sampled LFs with $M' \times N'$ SAIs from sparsely-sampled LFs with $M \times N$ SAIs. Moreover, we investigated the effect of the positions of SAIs involved in the sparsely-sampled LF input on the reconstruction quality via task $2 \times 2 - 8 \times 8$.

Both quantitative and qualitative results will be shown in the following subsections. Reconstruction quality is measured with PSNR and SSIM, averaged over all synthesised novel SAIs. Due to limited space, we only report the average result for all data entries in each dataset. The $(5, 5)$-th SAI of the reconstructed densely-sampled LF is chosen for display. Both training and testing codes are publicly available[2].

4.1 $3 \times 3 - 7 \times 7$ Light Field Reconstruction

For the task $3 \times 3 - 7 \times 7$, we compared with Kalantari *et al.* [13] and Wu *et al.* [14]. We set the number of spatial-angular alternating convolutional layers to 4. Comparisons were performed on the *30 Scenes* dataset [13], the *reflective-29* and *occlusion-16* LFs from the Stanford Lytro Lightfield Archive [28] and *Neurons* 20× from the Stanford Light Field microscope dataset [29]. The reconstruction

[2] https://github.com/angularsr/LightFieldAngularSR.

Table 2. Quantitative comparisons of the reconstruction quality of the proposed model and the state-of-the-art methods under the task $3 \times 3 - 7 \times 7$.

Algorithm	30 scenes	*Reflective-29*	*Occlusions-16*	*Neurons* 20×	Average
Wu *et al.* [14]	41.02/0.9875	46.10/0.9929	38.86/0.9852	29.34/0.9378	40.75/0.9861
Kalantari *et al.* [13]	43.73/0.9891	46.54/0.9953	37.97/0.9827	28.45/0.9274	43.18/0.9872
Ours 4L	**44.53/0.9900**	**47.85/0.9960**	**39.53/0.9873**	**30.69/0.9518**	**44.06/0.9889**

Table 3. Quantitative comparisons of reconstruction quality of the proposed model, Kalantari *et al.* and Wu *et al.* over *Buddha* and *Mona* from the HCI dataset.

Algorithm	*Buddha*	*Mona*	Average
Wu *et al.* [14]/SC	41.67/0.9975	42.39/0.9973	42.03/0.9974
Wu *et al.* [14]/SRCNN	41.50/0.9971	42.64/0.9976	42.07/0.9974
Wu *et al.* [14]	43.20/**0.9980**	44.37/**0.9982**	43.79/**0.9981**
Kalantari *et al.* [13]	42.73/0.9844	42.42/0.9858	42.58/0.9851
Ours 8L	**43.77**/0.9872	**45.67**/0.9920	**44.72**/0.9896

quality measured in PSNR and SSIM is shown in Table 2. For each LF, the results are the average of the luma component of all 40 novel SAIs. Our proposed model performs better for all datasets than the two comparing methods: with 0.88 dB and 3.31 dB reconstruction advantage over Kalantari *et al.* [13] and Wu *et al.* [14], respectively. A 2.3 dB advantage for the *Neurons* 20× dataset shows that the proposed LF reconstruction model generalizes well to different LF capturing devices.

4.2 $3 \times 3 - 9 \times 9$ Reconstruction on Large Disparity Light Field

To demonstrate that our model can work on LFs with larger disparities, the proposed model was modified for task $3 \times 3 - 9 \times 9$ and was trained with LFs from the HCI dataset [30], which are created with Blender software [31], with larger disparities compared with Lytro Illum captures. The LFs *Budda* and *Mona* are used for testing and the rest are used for training. For this task, we set the number of spatial-angular alternating convolution layers to 8. Due to limited number of training images, data augmentation was applied for obtaining more data training samples.

Comparison results with [14] are reported in Table 3. Using only 7 training LFs, our proposed method provides superior reconstruction quality on the luma component, averaged across all 72 novel SAIs.

4.3 $2 \times 2 - 8 \times 8$ Light Field Reconstruction

We carried out comparison with the method by Kalantari *et al.* [13] retrained with the same training dataset as ours. The method by Wu *et al.* [14] cannot be compared since their approach requires 3 views in each angular dimension

Table 4. Quantitative comparisons of reconstruction quality of the proposed model and Kalantari *et al.* under task $2 \times 2 - 8 \times 8$ over 222 real-world LFIs.

Algorithm	*30 Scenes*	*EPFL*	*Reflective*	*Occlusions*	Average
Kalantari *et al.* [13]	38.21/0.9736	38.70/0.9574	35.84/0.9416	31.81/0.8945	36.90/0.9452
Ours 16L	**39.22/0.9773**	**39.57/0.9637**	**36.47/0.9472**	**32.68/0.9061**	**37.76/0.9521**

to provide enough information for the bicubic interpolation step. Our testing dataset contains 30 test scenes from [13] (See footnote 1) and 118 LFs from the EPFL dataset [32][3] with diversified real-world scenes. To further evaluate the robustness of the algorithms, we also included the *Refractive and Reflective Surfaces* and the *Occlusions* categories from the Stanford Lytro Lightfield Archive [28], which contain 31 and 43 LFs, respectively. Note that the 8 LFs from the *Occlusions* category and 1 LF from the *Refractive and Reflective Surfaces* category were removed from testing as they were used for training. This test set contains 222 LFs which is sufficient to provide objective evaluation of model performance.

Reconstruction quality is measured with PSNR and SSIM averaged over the RGB channels, and over all 60 novel SAIs. As shown in Table 4, our proposed model with 16 alternating convolutions in the synthesis network obtains an average of 37.76 dB, 0.86 higher than that of Kalantari *et al.* [13].

Figure 4 further visually demonstrates that our algorithm is able to obtain better reconstruction quality compare with the state-of-the-art. As shown in the error maps, Kalantari *et al.* produces artifacts near the boundaries of the foreground objects. In most cases, thin edges cannot be reconstructed correctly, leaving blurred and overlapped regions between occluders and the background. Moreover, since our method fully explores the relationship among all SAIs in the reconstruction process, the LF structure is well preserved, leading to better EPI quality that can contribute to more accurate depth estimation.

4.4 $2 \times 2 - 8 \times 8$ Light Field Reconstruction with Extrapolation

Figures 5(a) and (b) show the average quality of each novel SAIs by Kalantari *et al.* [13] and the proposed approach under the task $2 \times 2 - 8 \times 8$, where it can be observed that reconstruction quality of the center SAIs has significantly worse quality compared with the novel SAIs near the input SAIs. The central view is furthest away from any of the input SAIs, therefore it poses greatest challenge to correctly infer the details. Based on this analysis, we investigated the possibility of combing interpolation and extrapolation for the LF reconstruction, which can make the average distances from all novel SAIs shorter to the input SAIs.

[3] https://jpeg.org/plenodb/lf/epfl/.

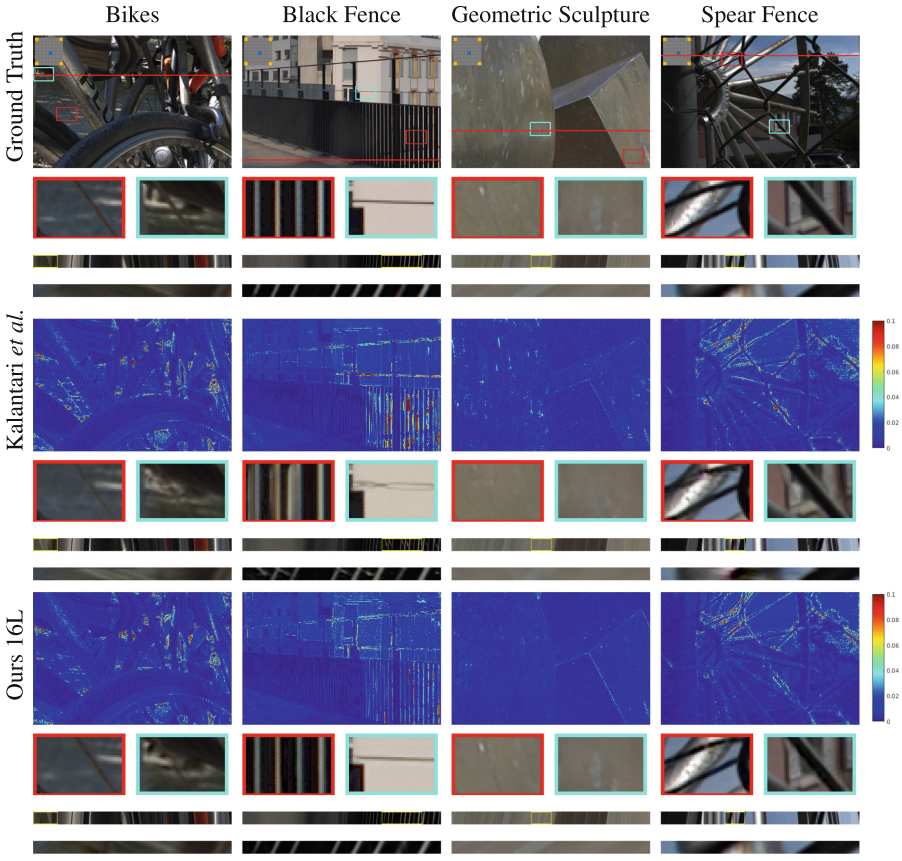


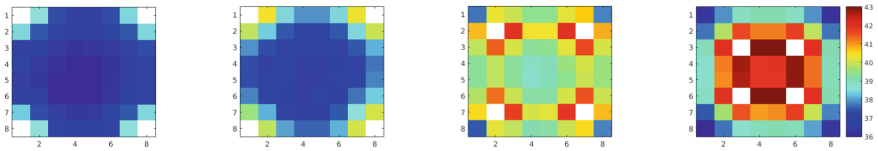
Fig. 4. Visual comparison of our proposed approach with Kalantari *et al.* [13] on the $(5, 5)$-th synthesised novel SAI for the task $2 \times 2 - 8 \times 8$. Selected regions have been zoomed on for better comparison. Digital zoom-in is recommended for more visual details.

We trained two models with the exact same network architecture as **Ours 8L**, however, with different input view position configurations as shown in Fig. 3(d) and (e), which we name as **Ours Extra. 1** and **Ours Extra. 2**, respectively. Note that for the first model, 1 row and column of SAIs are extrapolated while for the second model, 2 rows and columns of SAIs are extrapolated.

As shown in Table 5, when our model combines interpolation and extrapolation, an average of 2.5 dB improvement can be achieved for all novel SAIs on the 222 LFs dataset. Figures 5(c) and (d) also show the average quality of each novel SAIs by **Ours Extra. 1** and **Ours Extra. 2**, respectively. The significant gain in reconstruction quality indicates the potential for the proposed algorithm to be applied on LF compression [33,34].

Table 5. Quantitative comparisons of reconstruction quality of **Ours**, **Ours Extra. 1**, **Ours Extra. 2** and Kalantari *et al.* over 222 real-world LFs. For the proposed models, the number of spatial-angular alternating convolutions is set to 8.

Algorithm	*30 Scenes*	*EPFL*	*Reflective*	*Occlusions*	Average
Kalantari *et al.* [13]	38.21/0.9736	38.70/0.9574	35.84/0.9416	31.81/0.8945	36.90/0.9452
Ours	38.88/0.9750	39.29/0.9611	36.52/0.9466	32.58/0.9019	37.55/0.9495
Ours Extra. 1	40.79/0.9820	41.25/0.9705	**40.16/0.9667**	35.54/**0.9275**	39.93/**0.9632**
Ours Extra. 2	**40.93/0.9827**	**41.46/0.9717**	40.02/0.9651	**35.79**/0.9246	**40.09**/0.9631



(a) (b) (c) (d)

Fig. 5. Each subfigure displays the average reconstruction quality measured as PSNR at different SAI position under the task $2 \times 2 - 8 \times 8$ of different models. The white blocks indicate the input SAIs. From left to right: (a) Kalantari *et al.* [13], (b) **Ours**, (c) **Ours Extra. 1** and (d) **Ours Extra. 2**.

4.5 Depth Estimation

To verify that the densely-sampled LF generated from our proposed model not only produces high PSNR for each SAIs, but also well preserves the 3-D geometric structures among the SAIs, we further applied the depth estimation algorithm [3] on the reconstructed densely-sampled LF with 8×8 SAIs generated from a sparsely-sampled LF with 2×2 SAIs. Figure 6 shows in each row the depth maps based on the sparsely-sampled LFs, the densely-sampled LFs from Kalantari *et al.*, the densely-sampled LFs from our model and the ground-truth densely-sampled LFs. It can be observed that the depth maps from **Ours Extra. 1** are more accurate than those by Kalantari *et al.*

4.6 Runtime and Reconstruction Quality vs. Model Depth

The runtime and performance trade-off of our proposed model with different numbers of alternating convolutions are shown in Fig. 7. We can observe that the reconstruction quality by our model increases rapidly with the number of alternating convolutions increasing. Furthermore, the adoption of extrapolation leads to a significant improvement in reconstruction with a runtime of around 11 s, over 30× speed up compared with Kalantari *et al.* [13], on an Intel i7-6700K CPU @ 4.00 GHz without GPU acceleration. Moreover, the scalable structure in the synthesis network enables a trade-off between the reconstruction quality

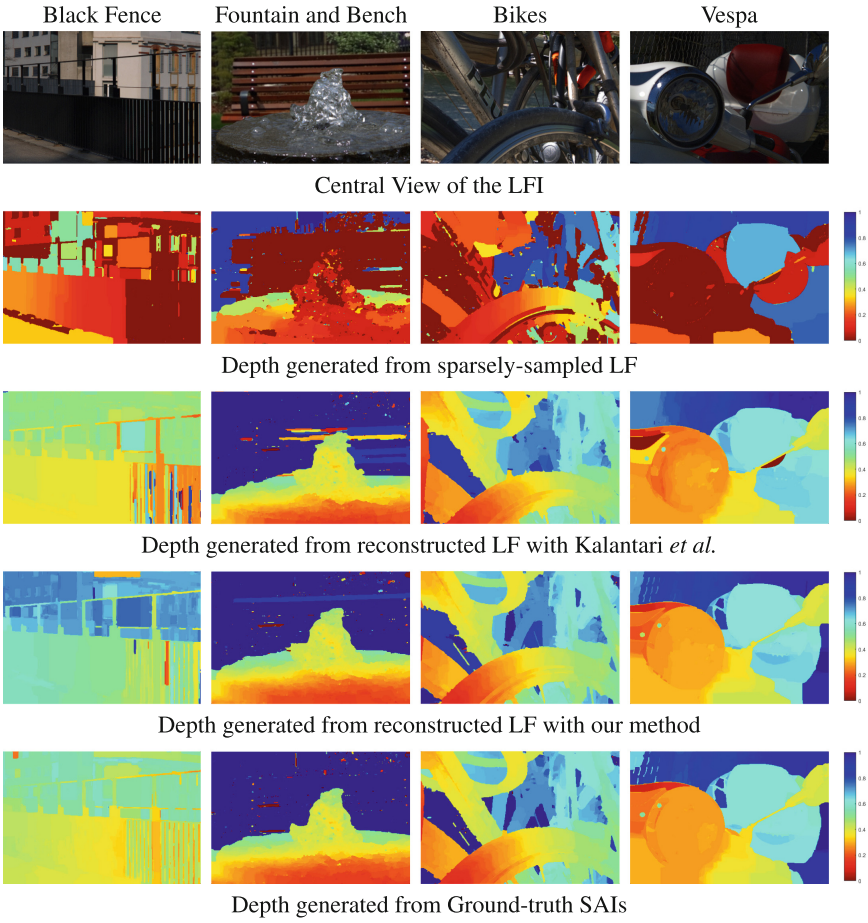
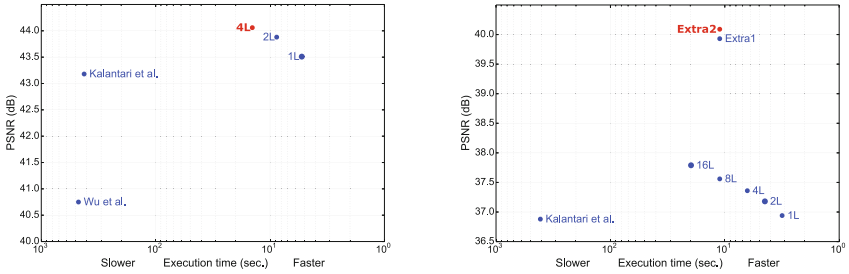


Fig. 6. Visual comparison of the depth estimation results from a sparsely-sampled LF, reconstructed densely-sampled LF from our proposed approach and Kalantari *et al.* [13] and a ground-truth densely-sampled LF.

and speed. For task $2 \times 2 - 8 \times 8$, our model with 16 alternating convolutions needs approximately 20 s. If speed is of priority, at similar reconstruction quality to Kalantari *et al.*, our model with 1 alternating convolution can provide over 130× speed up, taking only 3.15 s to process an LF.



(a) Task $3 \times 3 - 7 \times 7$ (b) Task $2 \times 2 - 8 \times 8$

Fig. 7. The trade-off between runtime and reconstruction quality at different model depth. Execution time in seconds were calculated as the average of 50 tests performed on an Intel i7-6700K CPU @ 4.00 GHz **without** GPU acceleration.

5 Conclusion and Future Work

We have presented a novel learning based framework for densely-sampled LF reconstruction. To characterize the high-dimensional spatial-angular clues within LF data accurately and efficiently, we have designed an end-to-end trained CNN that extensively employs spatial-angular alternating convolutions for fast feature transformation and stride-2 4-D convolutions for rapid angular dimension reduction. Moreover, our network synthesizes novel SAIs in a coarse-to-fine manner by first reconstructing a set of intermediate novel SAIs synthesized at the coarse angular dimension, then applying guided residual learning to refine the intermediate views at a finer level.

Extensive evaluations on real-world and synthetic LF scenes show that our proposed model is able to provide over 3 dB reconstruction quality in average than the state-of-the-art methods while being over 30× faster. Especially, our model can handle complex scenes with serious occlusions well. Moreover, our model is able to perform well under LFs with larger disparities, and more accurate depth can be inferred from the reconstructed densely-sampled LFs by our method. Considering the efficiency and effectiveness of the proposed CNN model in processing LF data, we believe such a design has great potential on LF compression, as well as a wide range of LF image processing tasks, including but not limited to LF spatial super-resolution, temporal super-resolution and depth estimation.

Acknowledgements. This work was supported in part by the CityU Start-up Grant for New Faculty under Grant 7200537/CS and in part by the Hong Kong RGC Early Career Scheme Funds 9048123 (CityU 21211518).

References

1. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 606–619 (2014)
2. Jeon, H.G., et al.: Accurate depth map estimation from a lenslet light field camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1547–1555 (2015)
3. Wang, T.C., Efros, A.A., Ramamoorthi, R.: Occlusion-aware depth estimation using light-field cameras. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3487–3495 (2015)
4. Chen, J., Hou, J., Ni, Y., Chau, L.P.: Accurate light field depth estimation with superpixel regularization over partially occluded regions. IEEE Trans. Image Process. **27**(10), 4889–4900 (2018)
5. Fiss, J., Curless, B., Szeliski, R.: Refocusing plenoptic images using depth-adaptive splatting. In: Proceedings of IEEE International Conference on Computational Photography, pp. 1–9 (2014)
6. 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)
7. Jones, A., McDowall, I., Yamada, H., Bolas, M., Debevec, P.: Rendering for an interactive 360 light field display. ACM Trans. Graph. **26**(3), 40 (2007)
8. Ihrke, I., Restrepo, J., Mignard-Debise, L.: Principles of light field imaging: briefly revisiting 25 years of research. IEEE Sig. Process. Mag. **33**(5), 59–69 (2016)
9. Wu, G., et al.: Light field image processing: an overview. IEEE J. Sel. Top. Sig. Process. **11**(7), 926–954 (2017)
10. Wilburn, B., et al.: High performance imaging using large camera arrays. ACM Trans. Graph. **24**(3), 765–776 (2005)
11. Lytro Illum. https://www.lytro.com/
12. Raytrix. https://www.raytrix.de/
13. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Trans. Graph. **35**(6), 193 (2016)
14. Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field reconstruction using deep convolutional network on EPI. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6319–6327 (2017)
15. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
16. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
17. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, no. 3, pp. 5835–5843 (2017)
18. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2790–2798 (2017)
19. Yoon, Y., Jeon, H.G., Yoo, D., Lee, J.Y., So Kweon, I.: Learning a deep convolutional network for light-field image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 24–32 (2015)

20. Yoon, Y., Jeon, H.G., Yoo, D., Lee, J.Y., Kweon, I.S.: Light-field image super-resolution using convolutional neural network. IEEE Sig. Process. Lett. **24**(6), 848–852 (2017)
21. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Comput. Sci. Tech. Rep. CSTR **2**(11), 1–11 (2005)
22. Rigamonti, R., Sironi, A., Lepetit, V., Fua, P.: Learning separable filters. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2754–2761 (2013)
23. Yan, L.Q., Mehta, S.U., Ramamoorthi, R., Durand, F.: Fast 4D sheared filtering for interactive rendering of distribution effects. ACM Trans. Graph. **35**(1), 7 (2015)
24. Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive separable convolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 261–270 (2017)
25. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A.A., Ramamoorthi, R.: A 4D light-field dataset and CNN architectures for material recognition. In: Proceedings of the European Conference on Computer Vision, pp. 121–138 (2016)
26. Vedaldi, A., Lenc, K.: MatConvNet - convolutional neural networks for MATLAB. In: Proceedings of the ACM International Conference on Multimedia, pp. 689–692 (2015)
27. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
28. Raj, A.S., Lowney, M., Shah, R., Wetzstein, G.: Stanford Lytro light field archive. http://lightfields.stanford.edu/
29. Levoy, M., Ng, R., Adams, A., Footer, M., Horowitz, M.: Light field microscopy. ACM Trans. Graph. **25**(3), 924–934 (2006)
30. Wanner, S., Meister, S., Goldluecke, B.: Datasets and benchmarks for densely sampled 4D light fields. In: VMV, pp. 225–226. Citeseer (2013)
31. Blender Online Community: Blender - a 3D modelling and rendering package. Blender Foundation, Blender Institute, Amsterdam
32. Rerabek, M., Ebrahimi, T.: New light field image dataset. In: Proceedings of the 8th International Conference on Quality of Multimedia Experience. Number EPFL-CONF-218363 (2016)
33. Hou, J., Chen, J., Chau, L.P.: Light field image compression based on bi-level view compensation with rate-distortion optimization. IEEE Trans. Circ. Syst. Video Technol. 1 (2018)
34. Chen, J., Hou, J., Chau, L.P.: Light field compression with disparity-guided sparse coding based on structural key views. IEEE Trans. Image Process. **27**(1), 314–324 (2018)