Light Field Reconstruction Using Deep Convolutional Network on EPI

Gaochang Wu[1,2], Mandan Zhao[2], Liangyong Wang[1], Qionghai Dai[2], Tianyou Chai[1], and Yebin Liu[2]

[1]State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China
[2]Department of Automation, Tsinghua University, Beijing, China
liuyebin@mail.tsinghua.edu.cn

Abstract

In this paper, we take advantage of the clear texture structure of the epipolar plane image (EPI) in the light field data and model the problem of light field reconstruction from a sparse set of views as a CNN-based angular detail restoration on EPI. We indicate that one of the main challenges in sparsely sampled light field reconstruction is the information asymmetry between the spatial and angular domain, where the detail portion in the angular domain is damaged by undersampling. To balance the spatial and angular information, the spatial high frequency components of an EPI is removed using EPI blur, before feeding to the network. Finally, a non-blind deblur operation is used to recover the spatial detail suppressed by the EPI blur. We evaluate our approach on several datasets including synthetic scenes, real-world scenes and challenging microscope light field data. We demonstrate the high performance and robustness of the proposed framework compared with the state-of-the-arts algorithms. We also show a further application for depth enhancement by using the reconstructed light field.

1. Introduction

Light field imaging [20, 13] is one of the most extensively used method for capturing the 3D appearance of a scene. Early light field cameras such as multi-camera arrays and light field gantries [33], required expensive custom-made hardware. In recent years, the introduction of commercial and industrial light field cameras such as Lytro [1] and RayTrix [2] have taken light field imaging into a new era. Unfortunately, due to restricted sensor resolution, they must make a trade-off between spatial and angular resolution.

To solve this problem, many studies have focused on novel view synthesis or angular super-resolution using a small set of views [25, 26, 28, 35, 37] with high spatial resolution. Recently, Kalantari *et al*. [16] proposed a learning-based approach to synthesize novel. views from a sparse set of views that performed better than other state-of-the-art approaches [14, 27, 29, 31, 36]. They employed two sequential convolutional neural networks (CNNs) to estimate the depth of the scene and predict the color of each pixel. Then, they trained the network by directly minimizing the error between the synthetic view and the ground truth image. However, due to the depth estimation-based method they introduced, their networks still resulted in artifacts such as tearing and ghosting, especially in the occluded regions and non-Lambertian surfaces. Fig. 1 shows the reconstruction results obtained by Kalantari *et al*. [16] and our proposed approach on the *Neurons* 20× case from the Stanford microscope light fields data [21]. The method by Kalantari *et al*. [16] results in blur in the occluded regions, while the proposed approach produces reasonable result

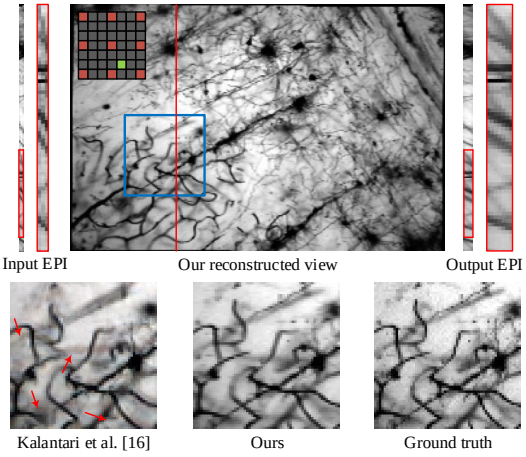

Figure 1. Comparison of light field reconstruction results on Stanford microscope light field data *Neurons* 20× [21] using 3 × 3 input views. The proposed learning-based EPI reconstruction produces better results in this challenging case.

even in this challenging case.

In this paper, we propose a novel learning-based framework to reconstruct high angular resolution light field from a sparse sample of views. One of our key insight is that the light field reconstruction can be modeled as learning-based angular detail restoration on the 2D EPI. Due to the special structure of the EPI, the learning-based reconstruction can be effectively implemented on it. Unlike the depth-based view synthesis approaches, the proposed method does not require depth estimation.

We further indicated (see Sec. 3) that the main problem in sparsely sampled light field reconstruction is the information asymmetry between the spatial and angular domain, where the high frequency portion in the angular domain is damaged by undersampling. This information asymmetry will cause ghosting effects when the light field is directly unsampled or super-resolved in the angular domain [24]. To suppress the ghosting effect caused by this information asymmetry and simultaneously take advantage of the spatial and angular information, we instead propose a "blur-restoration-deblur" framework on EPI. We first balance the information by removing the spatial high frequency information of the EPI. This step is implemented by convolving the EPI with a known blur kernel. We then apply a CNN to restore the angular detail of the EPI damaged by the undersampling. Finally, a non-blind deblur operation is used to restore the spatial detail suppressed by the EPI blur.

Extensive experiments on synthetic scenes and real-world scenes as well as microscope light field data validate that the proposed framework significantly improves the reconstruction in the occluded regions, non-Lambertian surfaces and transparent regions, and it produces novel views with higher numerical quality (4dB higher) compared to other state-of-the-art approaches. Moreover, we demonstrate that the reconstructed light field can be used to substantially enhance the depth estimation. The source code of our work will be made public.

2. Related Work

The main obstacle in light field imaging is the trade-off between spatial and angular resolution due to limited sensor resolution. Super-resolution techniques to improve spatial and angular resolution have been studied by many researchers [5, 6, 31, 35, 12]. In this paper, we mainly focus on approaches for improving the angular resolution of the light field. The related work is divided into two categories: those that use depth estimation and those that do not.

2.1. Depth image-based view synthesis

Wanner and Goldluecke [31] introduced a variational light field spatial and angular super-resolution framework by utilizing the estimated depth map to warp the input images to the novel views. They employed the structure tensor to obtain a fast and robust local disparity estimation. Based on Wanner and Goldluecke's work, a certainty map was proposed to enforce visibility constrains on the initial estimated depth map in [22]. Zhang *et al*. [37] proposed a phase-based approach for depth estimation and view synthesis. However, their method was specifically designed for a micro-baseline stereo pair, and causes artifacts in the occluded regions when extrapolating novel views. Zhang *et al*. [36] described a patch-based approach for various light field editing tasks. In their work, the input depth map is decomposed into different depth layers and presented to the user to achieve the editing goals. However, these depth image-based view synthesis approaches suffer when faced with occluded and textureless regions. In addition, they often focus on the quality of depth estimation, rather than the synthetic views.

In recent years, some studies for maximizing the quality of synthetic views have been presented that are based on CNNs. Flynn *et al*. [11] proposed a deep learning method to synthesize novel views using a sequence of images with wide baselines. Kalantari *et al*. [16] used two sequential convolutional neural networks to model depth and color estimation simultaneously by minimizing the error between synthetic views and ground truth images. However, in that study, the network is trained using a fixed sampling pattern, which makes it unsuitable for universal applications. In addition, the approach results in ghosting artifacts in the occluded regions and fails to handle some challenging cases.

In general, the depth image based view synthesis approaches [22, 31, 36, 37] use the estimated depth map to warp the input images to the novel views. In contrast, the learning-based approaches [11, 16] are designed to minimize the error between the synthetic views and the ground truth images rather than to optimize the depth map, resulting in better reconstruction results. However, these approaches still rely on the depth estimation; therefore, they always fail in occluded regions, non-Lambertian surfaces and transparent regions.

2.2. Light field reconstruction without depth

For sparsely sampled light fields, a reconstruction in Fourier domain has been investigated in some studies. Shi *et al*. [26] considered light field reconstruction as an optimization for sparsity in the continuous Fourier dimain. Their work sampled a small number of 1D viewpoint trajectories formed by a box and 2 diagonals to recover the full light field. However, this method requires the light field to be captured in a specific pattern, which limits its practical uses. Vagharshakyan *et al*. [28] utilized an adapted discrete shearlet transform to reconstruct the light field from a sparsely sampled light field in EPI space. However, they assumed that the densely sampled EPI was a square image, therefore, needed large number of input views. In addition,

the reconstruction exhibited poor quality in the border regions, resulting in a reduction of angular extent.

Recently, learning-based techniques have also been explored for the reconstruction without depth. Cho *et al*. [8] adopted a sparse-coding-based (SC) method to reconstruct light field using raw data. They generate image pairs using Barycentric interpolation. Yoon *et al*. [35] trained a neural network for spatial and angular super-resolution. However, the network used every two images to generate a novel view between them, thus it underused the potential of the full light field. Wang *et al*. [30] proposed several CNN architectures, one of which was developed for the EPI slices; however, the network is designed for material recognition, which is different with the EPI restoration task.

3. Problem Analysis and Formulation

For a 4D light field $L(x, y, s, t)$, where x and y are the spatial dimensions and s and t are the angular dimensions, a 2D slice can be acquired by gathering horizontal lines with fixed y^* along a constant camera coordinate t^*, denoted as $E_{y^*,t^*}(x, s)$. This 2D slice is called an epipolar plane image (EPI). Then, the low angular resolution EPI E_L is s downsampled version of the high angular resolution EPI E_H:

$$\mathbf{E}_L = \mathbf{E}_H \downarrow, \tag{1}$$

where $\downarrow$ denotes the down-sampling operation. Our task is to find an inverse operation F that can minimize the error between the reconstructed EPI and the original high angular resolution EPI:

$$\min_F ||\mathbf{E}_H - F(\mathbf{E}_L)||. \tag{2}$$

For a densely sampled light field, where the disparity between the neighboring views does not exceed 1 pixel, the angular sampling rate satisfies the Nyquist sampling criterion (the detail of this deduction can be found in [24]). One can reconstruct such a light field based on the plenoptic function; however, for light field sampled under the Nyquist sampling rate in the angular domain, the disparity is always larger than 1 pixel (see Fig. 2(a)). This undersampling of the light field destroys the high frequency detail in the angular domain, while the spatial information is complete. This information asymmetry between the angular and spatial information causes ghosting effect in the reconstructed light field if the angular resolution is directly upsampled (see Fig. 2(b)). The black line in the ground truth EPI (Fig. 2(e)) is continues, while the upsampled EPI (Fig. 2(b)) cannot reconstruct the line with large disparity. Note that this information asymmetry will always occur when the disparity between the neighboring views is larger than 1 pixel.

To ensure information symmetry between the spatial and angular information of the EPI, one can decrease the spatial resolution of the light field to an appropriate level. However, it is then difficult to recover the novel views with the original spatial quality, especially when a large downsampling rate has to be used in the case as shown in Fig. 2 (a). Rather than decreasing the spatial resolution of the light field, we extract the low frequency information by convolving the EPI with a 1D blur kernel in the spatial domain. Due to the coupling relationship between the spatial and angular domain [24], this step equals an anti-aliasing processing in the angular domain. Because the kernel is predesigned, the spatial detail can be easily recovered by using a non-blind deblur operation. Fig. 2(c) shows the blurred and upsampled result of the sparsely sampled EPI in Fig. 2(a). We now reformulate the reconstruction of EPI $\mathbf{E}_L$ as follows:

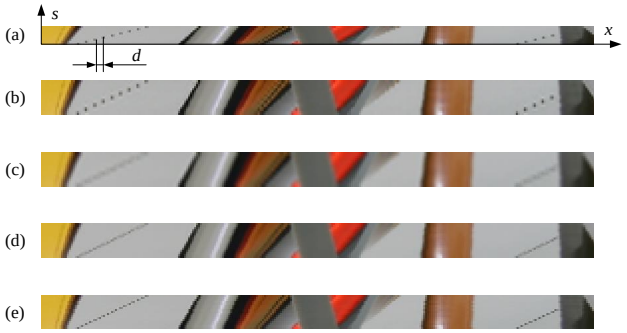

Figure 2. An illustration of EPI upsampling results. (a) The input low angular resolution EPI, where d is the disparity between the neighboring views (4 pixels); (b) The upsampling result using angular super-resolution directly cannot reconstruct an EPI with visual coherency; (c) The result after using EPI blur (on the spatial dimension) and bicubic interpolation (on the angular dimension); (d) The final high angular resolution EPI produced by the proposed algorithm; and (e) The ground truth EPI.

$$\min_f ||\mathbf{E}_H - D_\kappa f((\mathbf{E}_L * \kappa) \uparrow)||, \tag{3}$$

where $*$ is the convolution operator, κ is the blur kernel, $\uparrow$ is a bicubic interpolation operation that upsamples the EPI to the desired angular resolution, f represents an operation that recovers the high frequency detail in the angular domain, and D_κ is a non-blind deblur operator that uses the kernel κ to recover the spatial detail of the EPI suppressed by the EPI blur. In our paper, we model the operation f with a CNN to learn a mapping between the blurred low angular resolution EPI and the blurred high angular resolution EPI.

4. Proposed Framework

4.1. Overview

The EPI is the building block of a light field that contains both the angular and spatial information. We take advantage of this characteristic to model the reconstruction of the sparsely sampled light field as the learning-based angular information restoration on EPI (Eq. 3). An overview of our proposed framework is shown in Fig. 3.

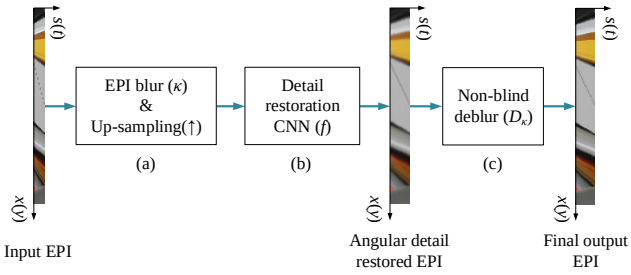

Figure 3. The proposed learning-based framework for light field reconstruction on EPI.

We first extract the spatial low frequency information of the EPI using EPI blur. Then we upsample it to the desired angular resolution using bicubic interpolation in the angular domain (see Fig. 3(a)). Then, we apply a CNN to restore the detail of the EPI in the angular domain (see Fig. 3(b)). The network architecture is similar to that in [9]. The main difference is that we apply a residual-learning method to predict only the angular detail of the EPI. The network detail is presented in Sec. 4.3. Finally, the spatial detail of the EPI is recovered through a non-blind deblur operation [18] (see Fig. 3(c)), and the output EPIs are applied to reconstruct the final high angular resolution light field. It should be noted that the CNN is trained to restore the angular detail that is damaged by the undersampling of the light field rather than the spatial detail suppressed by the EPI blur. An alternative approach is to model the deblur operation into the CNN; however, using that approach, the network will inevitably need to be deeper and will be slower to converge, making it more difficult to produce good results. Comparatively, the non-blind deblur is much more suitable to the task because the kernel is known.

To reconstruct the full light field using the sparsely sampled light field, the EPIs $E_{y^*,t^*}(x,s)$ and $E_{x^*,s^*}(y,t)$ from the input views are applied to reconstruct an intermediate light field. Then, EPIs from the novel views are used to generate the final light field.

4.2. Low frequency extraction based on EPI blur

To extract the low frequency of the EPI from only the spatial domain, we define the blur kernel in 1D space rather than defining a 2D image blur kernel. The following candidates are considered when extracting the low frequency part of the EPIs: the sinc function, the spatial representation of a Butterworth low pass filter of order 2 and the Gaussian function. The spatial representations of the filters are as follows:

$$\begin{aligned}\kappa_s(x) &= c_1 \mathrm{sinc}(x/(2|\sigma|)),\\ \kappa_b(x) &= c_2 e^{-|x/\sigma|}(\cos(|x/\sigma|) + \sin(|x/\sigma|)),\\ \kappa_g(x) &= c_3 e^{-x^2/(2\sigma^2)},\end{aligned} \tag{4}$$

where c_1, c_2 and c_3 are scale parameters, and σ is a shape parameter. In our paper, the kernels are discretized at the integer coordinate and limited to a finite window, i.e., $x \in [-4\sigma, 4\sigma]$. The kernel size is determined by the largest disparity (e.g., for the light field with largest disparity of 4 pixels, the shape parameter $\sigma = 1.5$, and the kernel size is 13). The scale parameters are used to normalize the kernels.

We evaluate these three kernels based on the following two principles: the final deblurred result must show visual coherency with the ground truth EPI, and the mean square error (MSE) between the blurred low angular resolution EPI and the blurred ground truth EPI is as minimal as possible:

$$\min_{\kappa} \frac{1}{n}\sum_{i=1}^{n} ||(\mathbf{E}_L^{(i)} * \kappa) \uparrow - \mathbf{E}^{(i)} * \kappa||^2, \tag{5}$$

where i is the index of the EPIs, n is the number of EPIs, $\mathbf{E}_L$ is the low angular resolution EPIs, and $\mathbf{E}$ is the ground truth high angular resolution EPIs. We evaluate the kernels on the Stanford Light Field Acheive [4], and the errors between the processed (blurred and upsampled) EPIs and the blurred ground truth EPIs are 0.153, 0.089 and 0.061 for the sinc, Butterworth and Gaussian kernels, respectively. The sinc function represents an ideal low pass filter in the spatial domain, and the low frequencies can pass through the filter without distortion. However, this ideal low pass filter causes ringing artifacts in the EPIs. The Butterworth kernel generates imperceptible ringing artifacts, while the Gaussian ensures that no ringing artifacts exist. Based on this observation and the numerical evaluation, the Gaussian function is selected to be the kernel for the EPI blur.

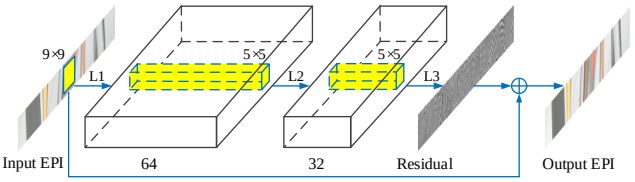

Figure 4. The proposed detail restoration network is composed of three layers. The first and the second layers are followed by a rectified linear unit (ReLU). The final output of the network is the sum of the predicted residual (detail) and the input.

4.3. Detail restoration based on CNN

For CNN based image restoration, Dong *et al*. [9] proposed a network for single image super-resolution named SRCNN, in which a high-resolution image is predicted from a given low-resolution image. Kim *et al*. [17] improved on that work by using a residual network with a deeper structure. Inspired by those pioneers, we design a residual network with three convolution layers to restore the angular detail of the EPIs.

4.3.1 CNN architecture

The architecture of the detail restoration network is outlined in Fig. 4. Consider an EPI that is convolved with the blur

kernel and up-sampled to the desired angular resolution, denoted as $\mathbf{E}'_L$ for short, the desired output EPI $f(\mathbf{E}'_L)$ is then the sum of the input $\mathbf{E}'_L$ and the predicted residual $\mathcal{R}(\mathbf{E}'_L)$:

$$f(\mathbf{E}'_L) = \mathbf{E}'_L + \mathcal{R}(\mathbf{E}'_L). \quad (6)$$

The network for the residual prediction comprises three convolution layers. The first layer contains 64 filters of size $1 \times 9 \times 9$, where each filter operates on 9×9 spatial region across 64 channels (feature maps) and used for feature extraction. The second layer contains 32 filters of size $64 \times 5 \times 5$ used for non-linear mapping. The last layer contains 1 filter of size $32 \times 5 \times 5$ used for detail reconstruction. Both the first and the second layers are followed by a rectified linear unit (ReLU). Due to the limited angular information of the light field used as the training dataset, we pad the data with zeros before every convolution operations to maintain the input and output at the same size.

We apply this residual learning method for the following reasons. First, the undersampling in the angular domain damages the high frequency portion (detail) of the EPIs; thus, only that detail needs to be restored. Second, extracting this detail prevents the network from having to consider the low frequency part, which would be a waste of time and result in less accuracy.

4.3.2 Training detail

The desired residuals are $\mathbf{R} = \mathbf{E}' - \mathbf{E}'_L$, where $\mathbf{E}'$ are the blurred ground truth EPIs and $\mathbf{E}'_L$ are the blurred and interpolated low angular resolution EPIs. Our goal is to minimize the mean squared error $\frac{1}{2}||\mathbf{E}' - f(\mathbf{E}'_L)||^2$. However, due to the residual network we use, the loss function is now formulated as follows:

$$L = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{R}^{(i)} - \mathcal{R}(\mathbf{E}'^{(i)}_L)||^2, \quad (7)$$

where n is the number of training EPIs. The output of the network $\mathcal{R}(\mathbf{E}'_L)$ represents the restored detail, which must be added back to the input EPI $\mathbf{E}'_L$ to obtain the final high angular resolution EPI $f(\mathbf{E}'_L)$.

We use the Stanford Light Field Archive [4] as the training data. The blurred ground truth EPIs are decomposed to sub-EPIs of size 17×17 with stride 14. To avoid overfitting, we adopted data augmentation techniques [10, 19] that include flipping, downsampling the spatial resolution of the light field as well as adding Gaussian noise. To avoid the limitations of a fixed angular up-sampling factor, we use a scale augmentation technique. Specifically, we downsample some EPIs with a small angular extent by factor 4 and the desired output EPIs by factor 2, then upsample them to the original resolution. The network is trained by using the datasets downsampled by both factor 2 and factor 4. We use the cascade of the network for the EPIs that are required to be up-sampled by factor 4. In practice, we extract more than 8×10^6 examples which is sufficient for the training. We select the mini-batches of size 64 as a trade-off between speed and convergence.

In the paper, we followed the conventional methods of image super-resolution to transform the EPIs into YCbCr space: only the Y channel (i.e., the luminance channel) is applied to the network. This is because the other two channels are blurrier than the Y channel and, thus, have less useful in the restoration [9].

To improve the convergence speed, we adjust the learning rate consistent with the increasing of the training iteration. The number of training iterations is 8×10^5 times. The learning rate is set to 0.01 initially and decreased by a factor of 10 every 0.25×10^5 iterations. When the training iterations are 5.0×10^5 , the learning rate is decreased to 0.0001 in two reduction steps. We initialize the filter weight of each layer using a Gaussian distribution with zero mean and standard deviation $1e^{-3}$. The momentum parameter is set to 0.9. Training takes approximately 12 hours on GPU GTX 960 (Intel CPU E3-1231 running at 3.40GHz with 32GB of memory). The training model is implemented using the *Caffe* package [15].

5. Experiment Results and Applications

In this section, we evaluate the proposed "blur-restoration-deblur" interpolation framework compared with the approach proposed by Kalantari *et al*. [16] and the typical depth-based approaches on several datasets including real-world scenes, microscope light field data and synthetic scenes. For the typical depth-based approaches, we first use current state-of-the-art approaches (Wang *et al*. [29], Jeon *et al*. [14]) to estimate the depth, then warp the input images to the novel view and blend by weighting the warped images [7]. We also evaluate each steps in the framework including: the performance without the "blur-deblur" steps; the residual-learning network by replacing the network with the SRCNN [9] and the sparse-coding-based method (SC) [34] in the detail restoration part. The quality of the synthetic views is measured by the PSNR against the ground truth image. In addition, we demonstrate how reconstructed light field can be applied to enhance the depth estimation[1].

5.1. Real-world scenes

We evaluate the proposed approach using 30 test scenes provided by Kalantari *et al*. [16] that were captured with a Lytro Illum camera ("30 scenes" for short) as well as two representative scenes, *Reflective* 29 and *Occlusion* 18, from the Stanford Lytro Light Field Achieve [3]. We use 3×3 views to reconstruct 7×7 light fields.

[1]More results of reconstructed light fields (figures and SSIM evaluation) and depth enhancement can be found in the supplementary file.

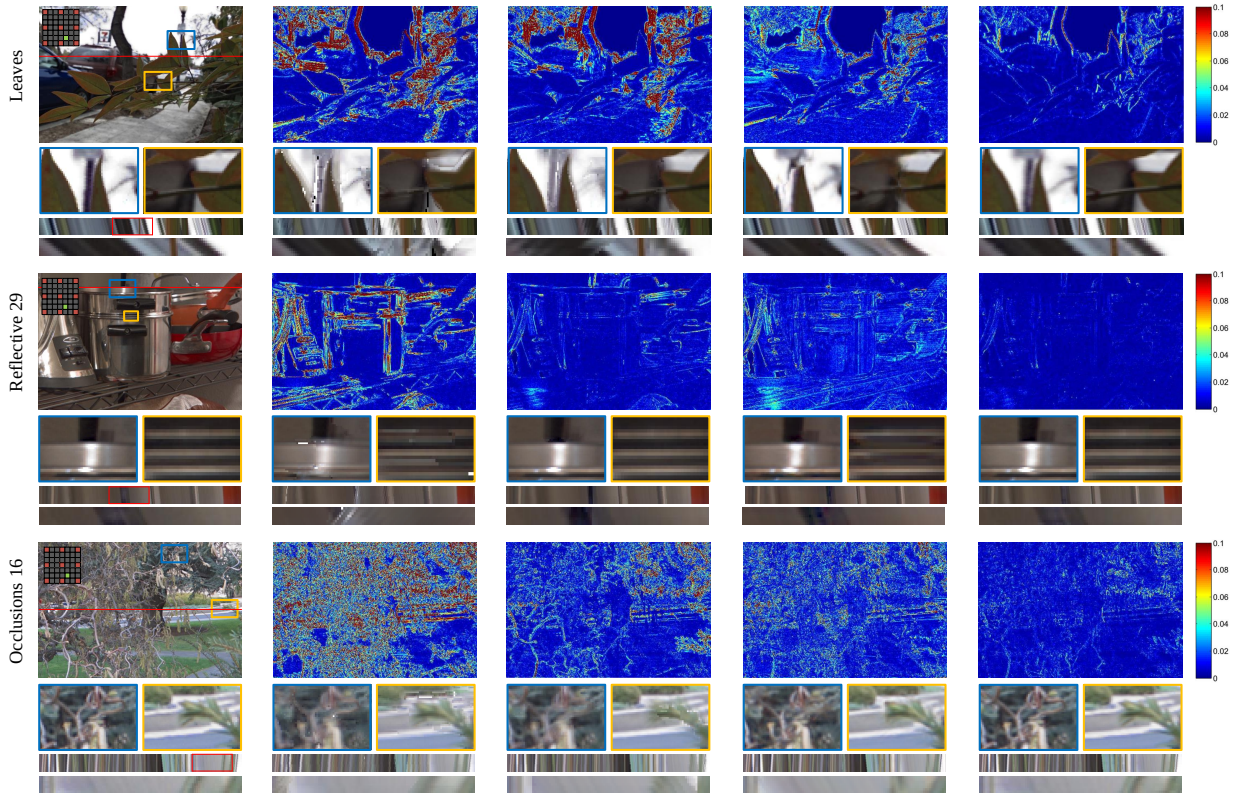

Figure 5. Comparison of the proposed approach against other methods on the real-world scenes. The results show the ground truth images, error maps of the synthetic results in the Y channel, close-up versions of the image portions in the blue and yellow boxes, and the EPIs located at the red line shown in the ground truth view. The EPIs are upsampled to an appropriate scale in the angular domain for better viewing. The lowest image in each block shows a close-up of the portion of the EPIs in the red box.

Table 1 lists the numerical results on the real-world datasets. The PSNR values are averaged over the 30 scenes. The CNNs in the approach by Kalantari *et al*. [16] are designed to minimizing the error between the synthetic views and the ground truth views. Therefore, they achieve better performance than other depth-based method among those common scenes. However, their networks were specifically trained for Lambertian regions, thus tend to fail in the reflective surface in the *Reflective* 29 case. Among these real-world scenes, our proposed framework is significantly better than other approaches. In addition, due to the information asymmetry, our proposed approach without the "blur-deblur" framework (denoted as "Ours/CNN only" in the table) produces lower quality light fields than those using the complete framework.

Fig. 5 depicts some of the results such as the *Leaves* from the 30 scenes, and *Reflective* 29 and *Occlusion* 16 scenes in the Stanford Lytro Light Field Achieve. The *Leaves* case includes some leaves with complex structure in front of a street. The case is challenging due to the overexposure of the sky and the occlusion around the leaves shown in the blue box. The results by Wang *et al*. [29] and Jeon *et al*. [14] show blurring artifacts around the leaves, and the result by Kalantari *et al*. [16] contains ghosting artifacts. The *Reflective* 29 case is a challenge scene because of the reflective surfaces of the pot and the kettle. The result by Wang shows blurring artifacts around the pot and the kettle. The approaches by Jeon *et al*. [14] and Kalantari *et al*. [16] produce better results, but the reconstructed light fields show discontinuities in terms of the EPIs. The *Occlusion* 16 case contains complicated occlusions that are challenging for view synthesis; consequently, their results are quite blurry around the occluded regions such as the branches and leaves. As demonstrated in the error maps and the close-up images of the results, the proposed approach achieves a high performance in terms of the visual coheren-

	30 scenes	*Reflective*29	*Occlusion*16
Wang *et al*. [29]	33.03	28.97	25.94
Jeon *et al*. [14]	34.42	40.27	32.10
Kalantari [16]	37.78	37.70	32.24
Ours/CNN only	37.15	44.84	35.89
Our proposed	**41.02**	**46.10**	**38.86**

Table 1. Quantitative results (PSNR) of reconstructed light fields on the real-world scenes [16, 3].

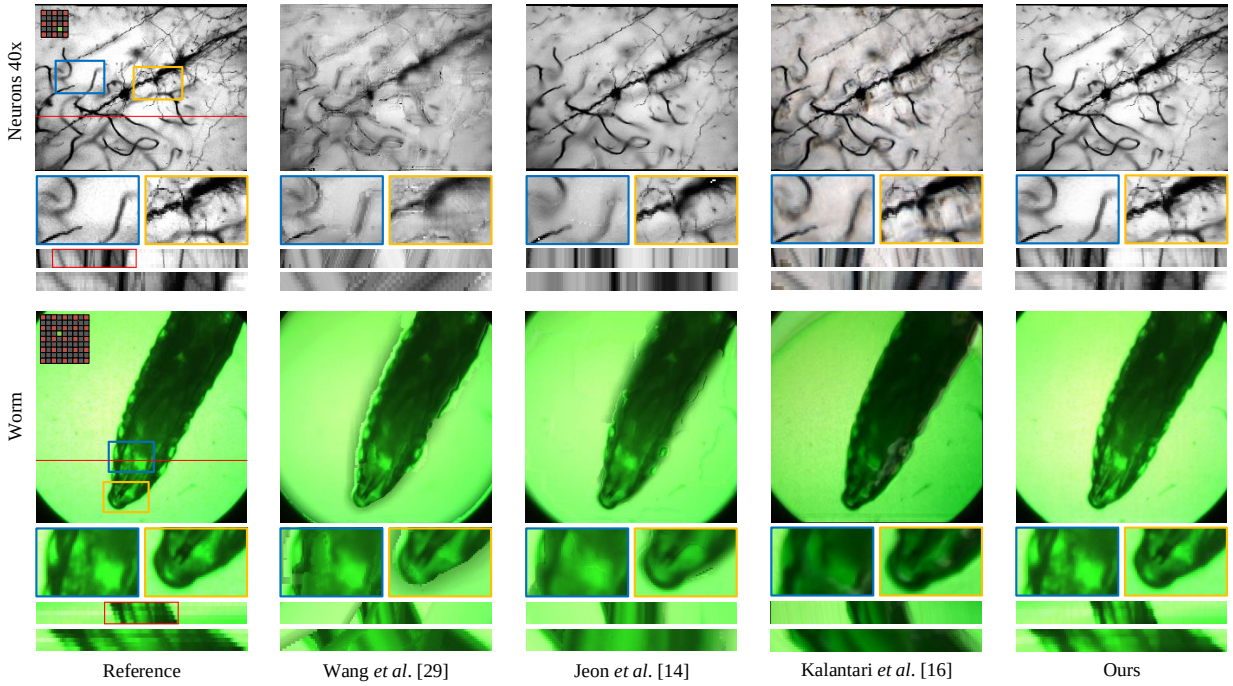

Figure 6. Comparison of the proposed approach against other methods on the microscope light field datasets. The results show the ground truth or reference images, synthetic results, close-up versions in the blue and yellow boxes, and the EPIs located at the red line shown in the ground truth view.

	Neurons 20×	*Neurons* 40×
Wang *et al*. [29]	17.45	13.21
Jeon *et al*. [14]	23.02	23.07
Kalantari *et al*. [16]	20.94	19.02
Our proposed	**29.34**	**32.47**

Table 2. Quantitative results (PSNR) of reconstructed light fields on microscope light field datasets [21].

cy of both the synthetic views and the EPIs.

5.2. Microscope light field dataset

In this subsection, the Stanford Light Field microscope datasets [21] and the camera array based light field microscope datasets provided by Lin *et al*. [23] are tested. These datasets include challenge light fields such as complicated occlusion relations and translucency. The numerical results are tabulated in Table 2, and the reconstructed views are shown in Fig. 6. We reconstruct 7×7 light fields using 3×3 views in the *Neurons* $40\times$ case, and 5×5 light fields using 3×3 views in the *Neurons* $40\times$ case. For the *Worm* case, 5×5 views are used to produce 9×9 light fields[2].

The *Neurons* $40\times$ case shows a Golgi-stained slice of rat brain, which contains complex occlusions. The result by Wang *et al*. [29] is quite blurry due to the errors in the estimated depth. Although the result by Jeon *et al*. [14] has a higher PSNR value, it fails to estimate the depth of the scene, which is visible in the EPI. The result produced by Kalantari *et al*. [16] has a higher quality in terms of the visual coherency. However, the result contains blurring and tearing artifacts in the occluded regions. Besides, the proposed approach shows denoising effect which can be seen in the close-up version. The *Worm* case is more simply structured but contains transparent objects such as the head of the worm. The depth-based approaches are not able to estimate accurate depth maps in those regions, which results in tearing and ghosting artifacts. Among these challenging cases, our approach produces plausible results in both the occluded and translucent regions.

[2]The quantitative evaluation is not performed on the *Worm* case because all the ground truth views are used as input. In the figure, we show a nearest view as the reference for the reconstructed view.

5.3. Synthetic scenes

We use the synthetic light field data from the HCI datasets [32] in which the spatial resolution is the same as the original inputs (768×768). The angular resolution of the output light field is set to 9×9 for comparison with the ground truth images, although we are able to produce light field of denser views. We use input light fields with different degrees of sparsity (3×3 and 5×5) to evaluate the performance of the proposed framework for different upsampling scale factors. Table 3 shows a quantitative evaluation of the proposed approach on the synthetic dataset compared with other methods. The approach by Kalantari *et al*. [16]

	Buddha		*Mona*	
Input	3×3	5×5	3×3	5×5
Wang *et al*. [29]	33.41	44.15	30.74	43.69
Jeon *et al*. [14]	41.19	44.06	40.95	42.67
Kalantari *et al*. [16]	34.05	34.51	32.53	32.59
Ours/SC [34]	41.67	41.79	42.39	44.40
Ours/SRCNN [9]	41.50	42.45	42.64	43.86
Our proposed	**43.20**	**46.42**	**44.37**	**51.07**

Table 3. Quantitative results (PSNR) of reconstructed light fields on the synthetic scenes of the HCI datasets [32]. The SC [34] and SRCNN [9] are applied to the proposed framework by replacing the proposed residual learning method and are denoted as ours/SC and ours/SRCNN, respectively.

produces lower quality than other depth-based approaches, because their CNNs are specifically trained on real-world scenes. The proposed approach achieves the highest PSNR values compared to the depth based methods. Moreover, the residual learning method produces better result than the SC and SRCNN approaches under the same framework.

5.4. Application for depth enhancement

In this section, we demonstrate that the proposed light field reconstruction framework can be used to enhance depth estimation. Table 4 gives the RMSE values of the depth estimation results from using the 3×3 inputs, the reconstructed 9×9 light fields produced by Kalantari *et al*. [16], our reconstructed 9×9 light fields and the ground truth 9×9 light fields on the HCI datasets [32]. Fig. 7 shows the depth estimation results on the *Cars*, *Reflective* 29, *Occlusion* 16, and the *Flowers and plants* 12 cases. We use the approach by Wang *et al*. [29] to estimate the depth of the scenes. The results show that our reconstructed light fields are able to produce more accurate depth maps that better preserve edge information than those produced by Kalantari *et al*. [16], e.g., the reflective surface of the red pan in the *Reflective* 29 and the branches in front of the left car in the *Cars*. Moreover, the enhanced depth maps are close to the ones produced by using the ground truth light fields.

6. Limitation and Discussion

The proposed framework uses EPI blur to extract the low frequency portion of the EPI in the spatial domain, where the size of the blur kernel is determined by the largest disparity between the input neighboring views. The non-blind deblur is not able to recover high quality EPIs when the kernel size is too large, and the maximum disparity we can handle is 5 pixels. For spatial aliasing input, our method cannot remove such artifacts but can give novel views with similar quality as those of the input. In addition, at least 3 views should be used in each angular dimension to provide enough information for the bicubic interpolation.

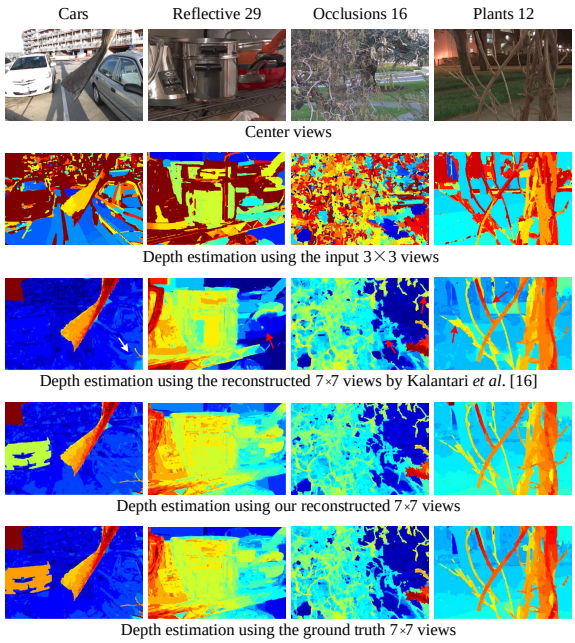

Figure 7. Depth estimation results using the reconstructed light fields. The arrows in the third row mark the depth errors caused by the artifacts of the reconstructed light fields.

	Buddha	*Mona*	*Horses*
Input 3×3 views	0.2926	0.2541	0.3757
Kalantari *et al*. [16]	0.1576	0.0829	0.1212
Ours	0.0401	0.0517	0.0426
GT light fields	0.0393	0.0529	0.0383

Table 4. RMSE values of the estimated depth using the approach by Wang *et al*. [29] on HCI datasets.

7. Conclusion

We have presented a novel learning-based framework for light field reconstruction on EPI. To avoid the ghosting effects caused by the information asymmetry, the spatial low frequency information of the EPI is extracted via EPI blur and used as input to the network to recover the angular detail. The non-blind deblur operation is used to restore the spatial detail that suppressed by the EPI blur. The experimental results demonstrate that the proposed framework outperforms state-of-the-art approaches in occluded and transparent regions and on non-Lambertian surfaces such as challenging microscope light field datasets.

Acknowledgments

This work was supported by the National key foundation for exploring scientific instrument No.2013YQ140517, the NSF of China grant No.61522111, No.61531014 and No.61673095.

References

[1] Lytro. https://www.lytro.com/.

[2] RayTrix. 3D light field camera technology. http://www.raytrix.de/.

[3] Stanford Lytro Light Field Archive. http://lightfields.stanford.edu/.

[4] Stanford (New) Light Field Archive. http://lightfield.stanford.edu/lfs.html.

[5] T. E. Bishop and P. Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE TPAMI*, 34(5):972–986, 2012.

[6] V. Boominathan, K. Mitra, and A. Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *ICCP*, pages 1–10. IEEE, 2014.

[7] G. Chaurasia, S. Duchêne, O. Sorkine-Hornung, and G. Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM TOG*, 32, 2013.

[8] D. Cho, M. Lee, S. Kim, and Y.-W. Tai. Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In *ICCV*, 2013.

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.

[10] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[11] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, 2015.

[12] X. Guo, Z. Yu, S. B. Kang, H. Lin, and J. Yu. Enhancing light fields through ray-space stitching. *IEEE TVCG*, 22(7):1852–1861, 2016.

[13] I. Ihrke, J. F. Restrepo, and L. Mignard-Debise. Principles of light field imaging: Briefly revisiting 25 years of research. *IEEE Signal Process. Mag.*, 33(5):59–69, 2016.

[14] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *CVPR*, pages 1547–1555, 2015.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678. ACM, 2014.

[16] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6), 2016.

[17] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2015.

[18] D. Krishnan and R. Fergus. Fast image deconvolution using hyper-laplacian priors. In *Advances in Neural Information Processing Systems*, pages 1033–1041, 2009.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[20] M. Levoy and P. Hanrahan. Light field rendering. In *Siggraph*, pages 31–42, 1996.

[21] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz. Light field microscopy. *ACM Transactions on Graphics (TOG)*, 25(3):924–934, 2006.

[22] J. Li, M. Lu, and Z.-N. Li. Continuous depth map reconstruction from light fields. *IEEE TIP*, 24(11):3257–3265, 2015.

[23] X. Lin, J. Wu, G. Zheng, and Q. Dai. Camera array based light field microscopy. *Biomedical Optics Express*, 6(9):3179–3189, 2015.

[24] Z. Lin and H.-Y. Shum. A geometric analysis of light field rendering. *International Journal of Computer Vision*, 58(2):121–138, 2004.

[25] S. Pujades, F. Devernay, and B. Goldluecke. Bayesian view synthesis and image-based rendering principles. In *CVPR*, pages 3906–3913, 2014.

[26] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand. Light field reconstruction using sparsity in the continuous fourier domain. *ACM TOG*, 34(1):12, 2014.

[27] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *ICCV*, pages 673–680, 2013.

[28] S. Vagharshakyan, R. Bregovic, and A. Gotchev. Image based rendering technique via sparse representation in shearlet domain. In *ICIP*, pages 1379–1383. IEEE, 2015.

[29] T.-C. Wang, A. A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *ICCV*, pages 3487–3495, 2015.

[30] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi. A 4d light-field dataset and cnn architectures for material recognition. In *ECCV*, pages 121–138. Springer, 2016.

[31] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE TPAMI*, 36(3):606–619, 2014.

[32] S. Wanner, S. Meister, and B. Goldlücke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modeling & Visualization*, pages 225–226, 2013.

[33] 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. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 765–776. ACM, 2005.

[34] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. pages 1–8, 2008.

[35] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *CVPRW*, pages 24–32, 2015.

[36] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu. Plenopatch: Patch-based plenoptic image manipulation. *IEEE TVCG*.

[37] Z. Zhang, Y. Liu, and Q. Dai. Light field from micro-baseline image pair. In *CVPR*, pages 3800–3809, 2015.