Revisiting Spatio-Angular Trade-off in Light Field Cameras and Extended Applications in Super-Resolution

Hao Zhu, Mantang Guo, Hongdong Li, Qing Wang, Antonio Robles-Kelly

Abstract—Light field cameras (LFCs) have received increasing attention due to their wide-spread applications. However, current LFCs suffer from the well-known *spatio-angular trade-off*, which is considered an inherent and fundamental limit for LFC designs. In this paper, by doing a detailed optical analysis of the sampling process in an LFC, we show that the effective resolution is generally higher than the number of micro-lenses. This contribution makes it theoretically possible to super-resolve a light field. Further optical analysis proves the "2D predictable series" nature of the 4D light field, which provides new insights for analyzing light field using series processing techniques. To model this nature, a specifically designed epipolar plane image (EPI) based CNN-LSTM network is proposed to super-resolve a light field in the spatial and angular dimensions simultaneously. Rather than leveraging semantic information, our network focuses on extracting geometric continuity in the EPI domain. This gives our method an improved generalization ability and makes it applicable to a wide range of previously unseen scenes. Experiments on both synthetic and real light fields demonstrate the improvements over state-of-the-arts, especially in large disparity areas.

Index Terms—Spatio-angular trade-off, Light field reconstruction, Super-resolution, Epipolar plane image, LSTM

1 Introduction

The light field camera [1], [2] is becoming more and more popular. Due to its capability to capture the whole 4D light field [3], [4] in a single shot, it enables new imaging features such as refocusing [5] and free-viewpoint roaming [6]. However, the performance of current LFCs is limited by the well-known *spatio-angular trade-off* [7], namely, the notion that the product between the spatial resolution and angular resolution must not exceed the sensor resolution.

Several methods have been proposed to recover a high angular resolution light field from a low-resolution input (Fig.1). However, there are still several challenges in current solutions. For depth-based methods [9], [10], [11], [12], [13], [14], the results are prone to errors in the depth estimation, which may cause artifacts on occlusion boundaries. Additionally, since each view is reconstructed independently, the geometric consistency between views can not be guaranteed.

Recently, learning-based light field reconstruction methods have also been explored. Kalantari et al. [15] proposed two convolutional neural networks (CNNs) to estimate the depth and predict colors sequentially. However, since an explicit depth map has to be estimated, their method is still prone to estimation error. Wu et al. [16], [17] tackled the issue with depth-based approaches by focusing on learning EPI super-resolution. They eliminated the information asymmetry [18] between the spatial and angular dimensions by applying a blur operation on the EPI. However, such a blur operation can not handle large disparity areas, where the continuous EPI lines become discrete points. In this case, the information asymmetry still exists after the blur operation. Moreover, the EPI consistency is lost during the super-resolution process, which leads to fine structures in the image to be lost or over-smoothed in the reconstructed views.

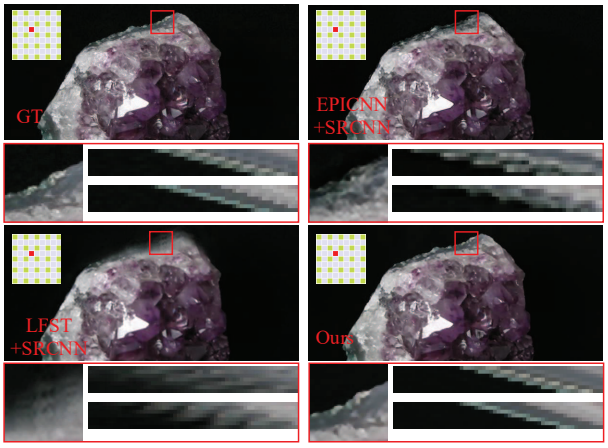


Fig. 1. A comparison of light field super-resolution results of the Amethyst [8]. Given a low-resolution (sparsely sampled) light field ($5 \times 5 \times 410 \times 307$), our method is able to produce a high-resolution (densely sampled) light field ($9 \times 9 \times 820 \times 614$). The bigger picture in each sub-figure shows the reconstructed view at $(4, 4)$ obtained by a number of different methods. In the bottom row of each sub-figure, the left panel shows a close-up image region (as indicated by the red box in the full image). On the right panel, we show the reconstructed horizontal and vertical EPIs.

In this paper, we first revisit the effective resolution of an LFC and find the "2D predictable series" feature of the 4D light field. Then a specifically designed learning-based

- *Hao Zhu, Mantang Guo and Qing Wang (corresponding author) are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, (e-mail: qwang@nwpu.edu.cn)*
- *Hongdong Li is with the Australian National University, ACT 0200, Australia.*
- *Antonio Robles-Kelly is with the Data61, CSIRO, ACT 2601, Australia.*
- *The work was supported by NSFC under Grant 61531014.*

method is proposed to super-resolve a light field in both angular and spatial dimensions.

One of our key insights is that the *spatio-angular trade-off* only holds when the LFC is in *generalized focused case* (Sec.3.1). In the defocused case, the effective spatial sampling rate can be higher than the number of micro-lenses in the conventional LFC [5]. Noting that the focused LFC [19] is not discussed here because the 4D light field can not be extracted from raw data directly. This insight is important since it provides a theoretical basis for further light field super-resolution beyond the resolution trade-off.

Secondly, the light field is proved to be a "2D predictable series" (Sec.3.2). The new pixel in the super-resolved sub-aperture image or the novel view is linearly connected with the pixels in the low-resolution light field and can be predicted. Thus the super-resolution for the 4D light field can be treated as a prediction problem for series data. This feature provides new insight for super-resolving light field using the series processing techniques.

Thirdly, a new light field dataset is proposed. Compared with previous publications, the proposed one provides three different resolutions for each light field, which is suitable for light field super-resolution tasks. Additionally, the provided high-accuracy depth maps may benefit the light field depth estimation and depth super-resolution tasks.

Last but not least, a learning-based framework for EPI super-resolution (Sec.4) is proposed. As summarized in Sec.3.2 that the light field is a "2D predictable series", the well-known convolutional long short term memory (LSTM) for series analysis is introduced for EPI super-resolution here. In contrast with previous super-resolution methods [15], [16], which leverage semantic information and content based inpainting, our network focuses on extracting and interpolating geometric continuity in the EPI. This gives our method a better generalization ability and makes it applicable to a wide range of previously unseen scenes. Experiments (Sec.5) on both synthetic and real light fields demonstrate the performance of the proposed LSTM layers and hint at significant improvements over state-of-the-art learning-based methods (>3dB), especially in large disparity areas.

2 Related Work

2.1 Light field sampling

Based on the two-parallel-plane (TPP) representation for light field sampling [3], [4], two types of LFCs were developed, namely, the conventional LFC [5] and the focused LFC [20]. However, they both suffer the *spatio-angular trade-off*. Bishop et al. [21] analyzed the optical path in the focused LFC. They pointed out that the aliasing effect in the spatial image contains new information, thus the resolution trade-off can be broken. The same conclusion was also summarized by Broxton et al. in [22] and Chang et al. in [23], where the diffraction effects are proved to be helpful for improving lateral resolution of the light field microscope using wave optics. Compared with [21], [22], [23], we focus on whole pixels instead of aliasing or diffraction and prove that multiple views in the conventional LFC record different point sets. The resolution of a light field can hence be improved by combining these point sets accordingly.

2.2 Depth-based methods

Light field reconstruction can be viewed as a special case of image based rendering, as the input and reconstructed novel views are all restricted in a 2D grid. So previous depth-based rendering techniques [9], [10], [11], [12], [13] can also be directly applied in light field reconstruction [14], [15], [24], [25]. However, there are two problems in the depth-based algorithms. Firstly, there are depth ambiguities in shadows, reflection and refraction areas where a correct depth may not be a good depth. Secondly, as each view is reconstructed independently, the view consistency may be broken in the reconstructed light field.

2.3 Non-depth-based methods

Considering the special grid characteristics of light field sampling, some signal processing cues have been used in light field reconstruction. These include, but are not limited to the dimension gap between a 3D focal stack and the 4D light field [26], the sparsity of light field sampling in continuous Fourier domain [27] and the sparse representation of EPI in shearlet transform domain [28].

Recently, CNNs have been used in light field reconstruction. Wu et al. [16], [17] tackled the light field reconstruction task viewing it as a one-dimensional EPI super-resolution problem and proposed a "blur-restoration-deblur" framework. Wang et al. [29] introduced a 4D CNN to directly super-resolve the 4D light field instead of the 2D EPI. Yeung et al. [30] explored the coarse characteristics of the sparsely-sampled light field and proposed the spatial-angular alternating convolutions to accelerate the reconstruction process. Guo et al. [31] explored very high angular light field reconstruction from multiple light fields using the residual network. All of these CNNs treat light field (or EPI) as a traditional 2D image, where each pixel is correlated with its standard square-like neighbouring system (4 or 8). However, note the neighbouring system size depends upon the direction of the EPI line in the light field and its displacement. Thus pixels with large disparities have a large neighbouring system. Wu et al. [32] improved the results using the fusion of sheared EPIs. However, this solution may also fail in occlusion areas with a large disparity. As a result, previous CNN-based methods work well for narrow baseline light fields while they often fail when applied to wide baseline light fields (see Fig.1).

2.4 Light field dataset

There have been several light field datasets in [8], [15], [33], [34], [35], [36], [37], [38], [39], [40], [41]. All these datasets have a fixed spatial resolution, which is an obstruction for further high-resolution light field reconstruction. Although a low-resolution light field can be generated by down-sampling each view image, it introduces prior in super-resolution process. Compared with previous datasets, the presented one has three advantages. Firstly, three different resolution light fields without any prior, which are all rendered using ray tracing techniques. Secondly, a larger occlusion ratio. The sampling of occlusion boundary areas is uneven and insufficient in light field [42], [43]. A dataset with large occlusion ratio can be better used to evaluate

the performance of light field applications. Thirdly, a larger disparity range. The so-called EPI line only exists in the light field with small disparity range (≤ 1 [14]) and it will become discrete points in large disparity area, where the successful technique in continuous areas may fail.

2.5 CNN and LSTM

CNN has become the hottest technique in almost all research areas of computer vision and has achieved significant improvement over previous techniques [44], [45], [46], [47]. Any input with particular patterns can be well modelled by CNN. LSTM [48] is a special recurrent neural network (RNN) which aims at modelling series data, such as text and speech. LSTM introduces 4 gates in each cell to solve the gradient vanishing and exploding problems in traditional RNN. Shi et al. [49] further improved the basic LSTM and proposed the convolutional LSTM for video analysis.

3 Optical Path Analysis in an LFC

In this section, we prove that the well-known *spatio-angular trade-off* only exists when an LFC is in the *generalized focused case*, *i.e.*, the disparities of all pixels in the recorded light field are integer values. In this case, all views in an LFC capture the same point set. Otherwise different views account for different point sets which are aliased with respect to other views. As a result, the effective spatial resolution of the conventional LFC is larger than the number of micro-lenses. In addition, the 4D light field is proved to be a 2D predictable series. The new pixel in the super-resolved light field, *i.e.*, both the new pixel in the high-resolution image or in the novel view, is linearly connected with the pixels in the low-resolution version and can be predicted. Therefore, the light field is naturally suitable to be processed by the techniques for series analysis.

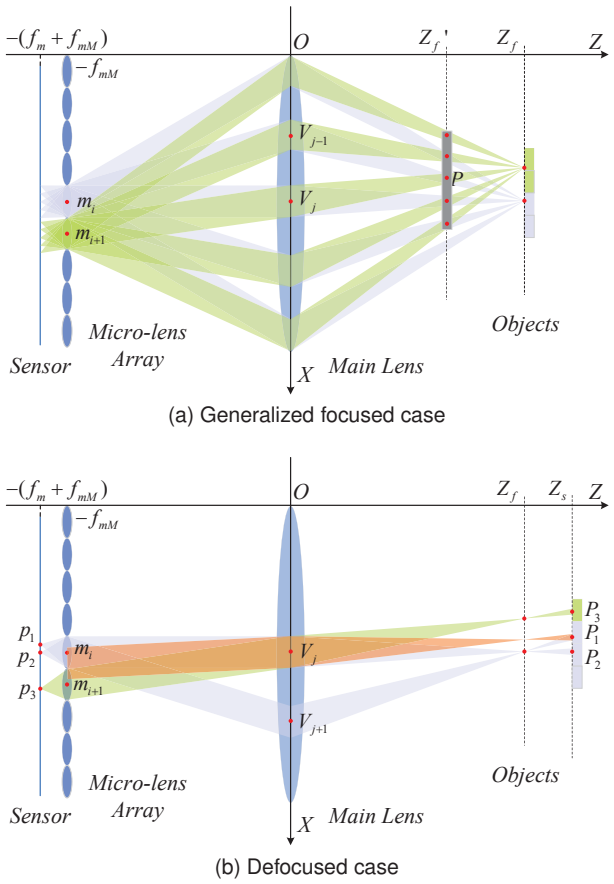


Fig. 2. Optical path in the conventional LFC. (a) In the generalized focused case, all views record a same 3D point set, thus the effective spatial resolution equals to the number of micro-lenses. (b) In the defocused case, different point sets are captured from different views due to the floating disparities, so the spatio-angular trade-off does not hold and the light field can be super-resolved.

3.1 On the number of recorded scene points

Generalized focused case

Fig.2a shows the optical path of an ideal conventional LFC, where all the pixels are covered by a micro-lens recording different views of the same point in the 3D space. In such a case, the depth Z_f of the scene point and the distance f_{mM} between the micro-lens array (MLA) and the main lens must meet the Gaussian imaging principle, *i.e.*, $\frac{1}{f_M} = \frac{1}{f_{mM}} + \frac{1}{Z_f}$, where f_M is the focal length of the main lens. Here, the *spatio-angular trade-off* holds and all recorded pixels are clear images of the objects at depth Z_f. That is, the recorded light field describes a consistent point set observed from different views.

If the scene depth varies, *i.e.*, the LFC is in the defocused case, the pixels covered by a micro-lens become a uniform sampling over a circular area in the 3D space (the gray areas in Fig.2a). In this case, different pixels under a micro-lens record different points in the 3D space. Note that the above trade-off also holds in some defocused situations. When the point in the 3D space is moved to the depth Z_f' in Fig.2a, the point P is only recorded once by the micro-lens m_{i+1} from the view V_j. Other views also record it at different positions, *e.g.*, the view V_{j-1} records it at the micro-lens m_i. In such a case, the images of point P from different angles are also recorded at different micro-lenses (boundary pixels are ignored here). In other words, the recorded light field is still a multi-view description of a same point set.

The above defocused case is similar to the focused case in the sense that different views in the recorded light field describe the same set of scene points. We call both the focused and the defocused at integer disparities cases *generalized focused case*, where previous *spatio-angular trade-off* holds.

Defocused case

Except for the above *generalized focused case*, different views of the recorded light field generally depict different scene point sets. As a result, the actual number of captured scene points is larger than the number of micro-lenses. Roughly speaking, the "resolution-trade-off" is broken in this case. Fig.2b illustrates the defocused case. Pixels p_1 and p_2 under the micro-lens m_i record two different points P_1 and P_2 from the views V_{j+1} and V_j, respectively. Note that the ray passing through the point P_1 from the view V_j to the MLA (the orange areas in Fig.2b) is "aliased" by the micro-lenses m_i and m_{i+1}. We can also trace the ray for the micro-lens m_{i+1} from the view V_j to the space point P_3. It can be seen that P_1 drops between the points P_2 and P_3. Because P_2 and P_3 are the nearest points in the view V_j, the point P_1, which is recorded by the view V_{j+1}, is not recorded by the

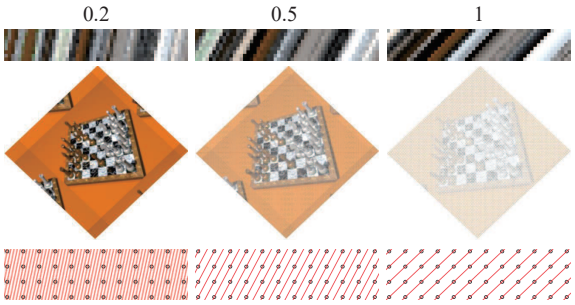


Fig. 3. The number of recorded pixels changes when the baseline in light field sampling changes. From top to bottom we show the EPIs, the reconstructed point clouds and the sketch maps of light field sampling. From left to right we show the light fields with 0.2, 0.5 and 1 pixel disparity. Note that the light field with 0.2 pixel disparity records the larger number of points.

view V_j. Thus the number of effectively recorded points in a conventional LFC is larger than the resolution of the view V_j. Since the resolution of each view equals the number of micro-lenses in the conventional LFC, the effective sampling resolution becomes larger than the number of micro-lenses.

We also provide an intuitive explanation of the above analysis on the EPI. Fig.3 shows EPIs and the corresponding point clouds under different disparity levels. Three light fields are captured with different baselines. It is noticed that the number of recorded points is different in these light fields and the one with 0.2 disparity records the most points. The sketches in the third row of Fig.3 reveal the reason well. When the disparity is 0.2, it can be seen that the red line passes through an entire pixel once every 5 views; in other words, views $\{1, 5, 9, ...\}$ sample same point set while views $\{2, 3, 4, 6, ...\}$ sample other point sets. When the disparity equals to 1, all views sample the same set of points. Thus, the light field with 0.2 disparity records the largest number of scene points. In summary:

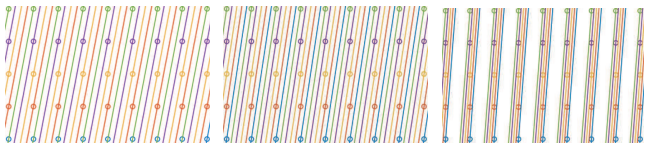
Proposition 1. *The* spatio-angular trade-off *in LFCs only holds when the LFC is in the* generalized focused case. *This is due to the fact that the depth has a continuous and complex distribution in a real-world scene. Thus, the effective spatial resolution of the conventional LFC is larger than the number of micro-lenses. In such a case, the light field can be super-resolved.*

Maximum super-resolution ratio under regular up-sampling

Given an LFC that contains M micro-lenses where each micro-lens covers N pixels, the maximum super-resolution ratio under regular up-sampling can be derived by analyzing the disparity distribution of the sampled points.

For the simplest case that the LFC is in the generalized focused case, the maximum super-resolution ratio for such a light field is only 1. Because all pixels have integer disparities, all views record the same point set and there are no performance gains for light field super-resolution compared with the single image super-resolution task.

For the defocused case, the maximum super-resolution ratio K is obtained from the following equation.



(a) $d(p) = 1/4$ pixel (b) $d(p) = 1/5$ pixel (c) $d(p) = 1/10$ pixel

Fig. 4. Demonstration of the super-resolution for different disparities.

$$\forall p,\ S_{d(p)} = \left(\bigcup_{t=1}^{N-1} \{td(p) - \lfloor td(p) \rfloor\} \right)$$

$$K^* = \left\{ \hat{K} \mid \exists \hat{K} \in \bigcup_{t=2}^{N} \{t\},\ S = \bigcup_{t=1}^{\hat{K}-1} \left\{ \frac{t}{\hat{K}} \right\},\ \text{s.t. } S_{d(p)} \subseteq S \right\} \quad (1)$$

$$K = \begin{cases} 1 & K^* = \emptyset \\ \max K^* & K^* \neq \emptyset \end{cases}$$

where $d(p)$ refers to the disparity of pixel p, $\lfloor \rfloor$ is rounding down operation, $\max K^*$ returns the maximum element of the set K^*. For the sake of simplicity, occlusion and boundary pixels are ignored here. All pixels are firstly projected to a fixed view, *i.e.*, the first row in the above equation. $S_{d(p)}$ records the positions of the fraction part of all projected pixels. The set K^* records all possible super-resolution ratios. It is noticed that not all floating disparities can benefit the super-resolution. Fig.4 demonstrates the comparison of three light fields with disparities $1/4$, $1/5$ and $1/10$ pixel. Because the number of views is 5, the super-resolution ratios are 4, 5 and 2 when $d(p) = 1/4$, $1/5$ and $1/10$, respectively. Finally, the maximum value in all possible ratios is selected as the upper bound of the super-resolution. Noting that, for other floating values without reasonable sense, the maximum super-resolution ratio is not discussed due to the irregular sampling.

3.2 Light field as 2D predictable series

Given an LFC with a low-resolution (Fig.5a), an "inverse ray tracing" operation is firstly applied to find the 3D positions P_1, P_2, P_3, P_4 of all recorded pixels. With the accurate positions of these 3D points, their imaging pixel positions in any views or imaging view positions in any pixels can be predicted by a "ray tracing" process (Fig.5b). For example, the imaging pixel m_* of the point P_2 in the view V_k can be derived by,

$$m_* = f_{mM}\left(\frac{1}{Z_1} - \frac{1}{Z_f}\right)(V_l - V_k) + m_j. \quad (2)$$

In other words, the light field can be seen as a "predictable" series in the angular dimension. Additionally, when fixing the pixel position m_j, the point P_3 will be observed by the view V_* where

$$V_* = V_l - (m_j - m_i)\frac{1}{f_{mM}\left(\frac{1}{Z_f} - \frac{1}{Z_2}\right)}. \quad (3)$$

Here, the light field is also a "predictable" series in the spatial dimension. In summary, the light field can be seen as a 2D predictable series.

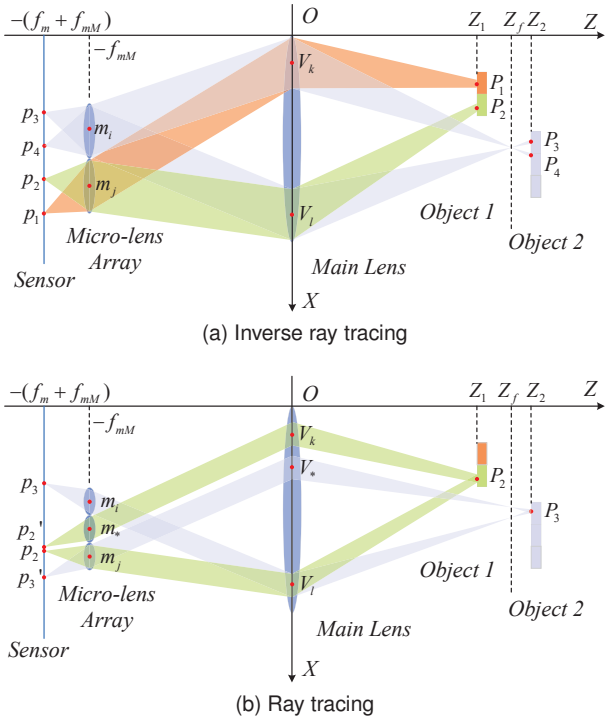


Fig. 5. Optical path of the light field super-resolution process in the conventional LFC.

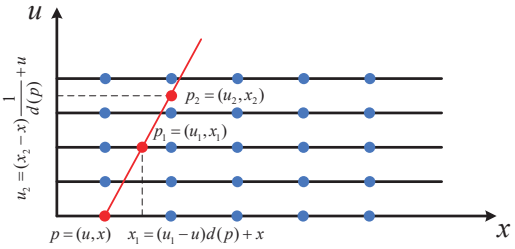


Fig. 6. For each EPI line, it can be projected to angular or spatial axis when fixing one of the axes for $d \in (0, \infty)$.

To better understand the above "predictable" feature, an intuitive demonstration on the EPI is provided in Fig.6. For each light ray $p = (u, x)$ in free space (without occlusion), there is a corresponding ray $p_1 = (u_1, x_1)$ describing the same 3D point in any other view u_1, such that

$$x_1 = (u_1 - u)d(p) + x, \tag{4}$$

where $d(p)$ refers to the disparity of p. In such a case, the light field is a predictable series in the angular dimension.

From another point of view, there is a corresponding ray $p_2 = (u_2, x_2)$ in any other pixel position x_2, such that

$$u_2 = (x_2 - x)\frac{1}{d(p)} + u. \tag{5}$$

Therefore, the light field is also a predictable series in the spatial dimension. In summary:

Proposition 2. *The 4D light field can be seen as a 2D predictable series. Any new pixel in the high-resolution sub-aperture image or the novel view is linearly connected with the pixels in the low-resolution light field and can be predicted.*

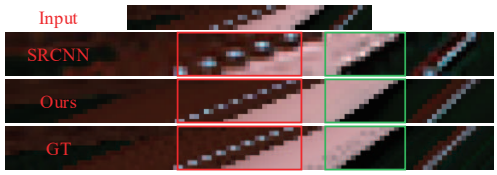


Fig. 7. Different types of continuity. From top to bottom, we show the input low-resolution EPI, super-resolved EPI from [51], ours and ground truth, respectively. Previous image super-resolution CNN copes well with "continuous continuity" (green boxes), failing for "discontinuous continuity" (red boxes).

4 EPI Super-resolution using a CNN-LSTM Network

While the above section shows it is possible to recover high-resolution light field defying the conventional spatio-angular trade-off, this is still not a straightforward task. The main difficulty strives from how to super-resolve a light field while keeping the consistency across different views. Most existing light field super-resolution methods are either based on depth recovery [15], [24], [50] or based on EPI analysis [16], [29], [30]. The former approaches are overly sensitive to errors in depth estimation, often failing to maintain cross-view consistency. The latter ones treat EPIs as a regular digital image, failing to capture the EPI nature of continuous traces corresponding to pixels across multiple views. In this section, we first discuss the issue of continuity preservation in light field super-resolution. Then we propose a novel CNN-LSTM architecture tailored for EPI super-resolution.

4.1 Different continuities

There are two types of continuities in a light field, *i.e.*, the "continuous continuity" and "jumping continuity" (Fig.7). When the disparity is small, EPI lines are continuous (green boxes) such that previous single image super-resolution methods [51], [52], [53] can be applied directly in such a case. However, the continuous EPI lines become discrete points (red boxes) when the disparity increases. Previous image super-resolution methods treat the discrete EPI line as independent points, so EPI line may be lost or over-smoothed in the super-resolved new views, leading thin structure objects missing or becoming unclear. Compared with "continuous continuity", the "jumping continuity" is more common in light field reconstruction. Because novel views can be synthesized directly by interpolating in 4D space [3], [54], [55] when the disparity is small, it is unnecessary to use expensive reconstruction techniques.

It is hard to super-resolve the EPI lines meeting the "jumping continuity" if treated EPI as a common 2D image. However, when treating EPI as a "2D predictable series" and super-resolving it using techniques for series modelling, the "jumping continuity" can be well handled (see Fig.7).

4.2 CNN-LSTM for EPI super-resolution

Given a high-resolution light field $LF_H(u, v, x, y)$ and its low-resolution version $LF_L(u, v, x, y)$, where (u, v) and (x, y) refer to the angular and spatial positions of a light ray

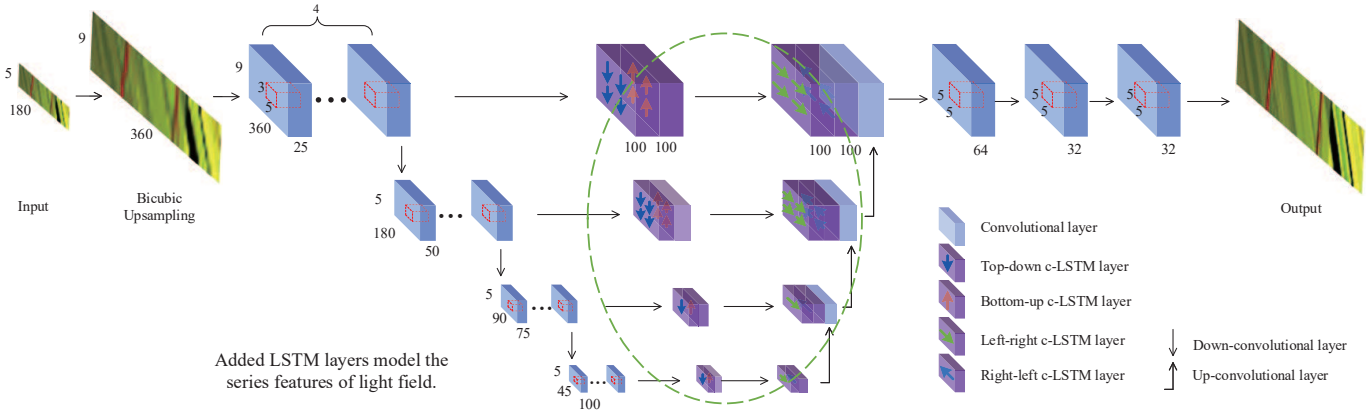


Fig. 8. The architecture of our neural network.

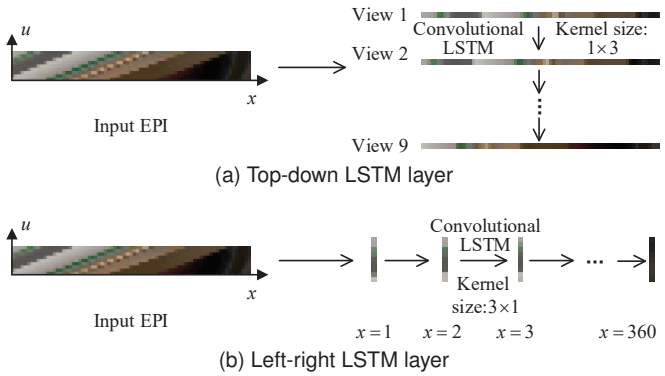


Fig. 9. The architectures of the convolutional LSTM layers for modelling EPI in the top-down and left-right directions.

respectively. The light field super-resolution task is to find an inverse function $\mathcal{F}$ that minimizes the error between the super-resolved light field and the original high-resolution light field,

$$\min_{\mathcal{F}} ||LF_H - \mathcal{F}(LF_L)||. \tag{6}$$

However, it is difficult to build a neural network to directly process the 4D data. We focus on the 2D slice super-resolution of the 4D expression, *i.e.*,

$$\min_{\mathcal{F}} ||LF_{H,(v^*,y^*)} - \mathcal{F}(LF_{L,(v^*,y^*)})|| \tag{7}$$

where $LF_{H,(v^*,y^*)}$ and $LF_{L,(v^*,y*)}$ refer to the 2D EPI slice of the high and low-resolution light fields when fixing (v, y) to (v^*, y^*), respectively.

Considering large disparities in the light field, we propose a CNN-LSTM network whose architecture is shown in Fig.8. The overall network is inspired by the U-network in EPI analysis [40], [56]. Our network has four "levels", where each of these accounts for the EPI at different resolutions. In contrast with previous work, four convolutional LSTM [49] layers are added at each level (the purple blocks in Fig.8) to model the series nature of the EPI in the top-down, bottom-up, left-right and right-left directions, respectively.

In this network, each convolutional LSTM has 100 channels. Fig.9 gives an illustration of the LSTM layers in top-down and left-right directions, respectively. For the top-down case, given a feature map $L^j(u, x)$ output from previous CNN layer j, the current top-down convolutional LSTM layer analyzes $L^j(u, x)$ and outputs the feature $L^{j+1}(u, x)$. The $L^j(u, x)$ is firstly separated into a sequence $L^j(u_1, x)$, $L^j(u_2, x)$, ..., $L^j(u_9, x)$. Then the convolutional LSTM function [49] is applied,

$$\begin{aligned}
i_t &= \sigma\left(W_{xi} * L^j(u_t, x) + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \circ \mathcal{C}_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf} * L^j(u_t, x) + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \circ \mathcal{C}_{t-1} + b_f\right) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh\left(W_{xc} * L^j(u_t, x) + W_{hc} * \mathcal{H}_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo} * L^j(u_t, x) + W_{ho} * \mathcal{H}_{t-1} + W_{co} \circ \mathcal{C}_t + b_o\right) \\
\mathcal{H}_t &= o_t \circ \tanh\left(\mathcal{C}_t\right),
\end{aligned} \tag{8}$$

where i_t,f_t,$\mathcal{C}_t$,o_t and $\mathcal{H}_t$ are the input gate, forget gate, cell output, output gate and hidden gate respectively. $*$ and $\circ$ are the convolution operator and Hadamard product, respectively. W_{xi}, W_{hi}, W_{xf}, W_{hf}, W_{xc}, W_{hc}, W_{xo} and W_{ho} are convolutional kernels. The size of such kernels is 1×3 when the LSTM direction is top-down or bottom-up, otherwise the size is 3×1. Finally, the output $L^{j+1}(u, x)$ of the current layer is,

$$L^{j+1}(u_t, x) = \mathcal{H}_t. \tag{9}$$

For the left-right case, a new input sequence is constructed by $L^j(u, x_1)$, $L^j(u, x_2)$, ..., $L^j(u, x_{180})$. Then the convolutional LSTM function is applied as shown in Eqns.8 and 9. For the bottom-up and right-left cases, inversing the sequences and following the above steps.

When processing a low-resolution input EPI, this is firstly scaled 2 times up in both angular and spatial dimensions using the bicubic interpolation. Before LSTM analysis at each level, 4 convolutional layers are applied. These layers have kernels of size 3×5. The channels of these convolutional kernels equal to $25 \times i$ for the i-th level. After LSTM analysis, three convolutional layers are added with kernel size 5×5 and channels 64, 32 and 32. Note that each convolutional layer is followed by a ReLU layer [57]. Different levels in Fig.8 are connected by down and up-convolutional layers with kernel size 3×3.

5 Experimental Results

We compare our method with a combination of state-of-the-art light field reconstruction methods such as EPICNN [16],

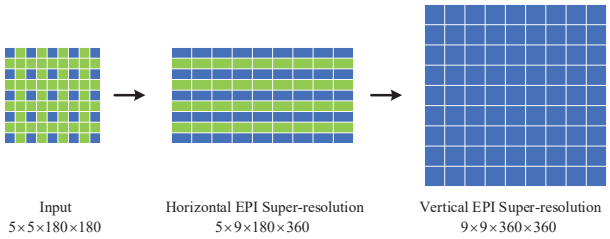


Fig. 10. Two steps of the proposed EPI based light field super-resolution method. Blue boxes refer to the input or reconstructed views, where the size of each box shows the resolution of each view image.

TABLE 1
Quantitative comparisons of different light field datasets (The occlusion ratio and disparity of our data are counted from light fields with the resolution 180×180).

	HCI [33]	Konstanz [38]	OurLFs
Number of light fields	7	24	100
Angular Resolution	9×9	9×9	9×9
Spatial Resolution	768×768	512×512	180×180 360×360 720×720
Occ. Ratio (mean)	0.59%	1.89%	6.07%
Occ. Ratio (max)	0.92%	6.64%	26.63%
Disparity (mean)	3.72	3.70	4.28
Disparity (max)	7.39	7.00	16.12

[17] and LFST [28], and image super-resolution methods such as SRCNN [51] and LFNet [58]. All the results shown here are evaluated using the code released by the authors.

We evaluate the performance of the proposed method both on synthetic and real light fields. All the quantitative comparisons shown here are the average values of all views. To train the network, we have rendered a synthetic dataset with 100 light fields (abbreviated as OurLFs) using the POV-Ray [59], [60] (see Sec.5.1). As our network is trained on the proposed synthetic data, to be fair, the synthetic data is only used to validate the efficacy of the LSTM layers and the number of levels. Real-world light fields from camera array [8], [35] are used for comparing with other methods. Here we have not used the light fields from the Lytro Illum camera due to its small disparity range. Because the proposed method works on EPI representation, it needs two super-resolution operations to recover a high-resolution light field from a low-resolution input (see Fig.10).

5.1 Datasets

In order to train and evaluate our network, we build an automatic light field generator (ALFG) based on POV-ray to render 100 light fields. Fig.11 shows some examples. For training and testing, we have included various challenging environments in our dataset. These include inter-reflection, occlusion, shadowing, various illumination conditions and structures with fine detail. Tab.1 shows a quantitative comparison between the presented and previous datasets [33], [38]. Noting that, because the ground truth depth maps of testing light fields in [38] are not publicly available, only 24 light fields are counted in Tab.1. Additionally, the dataset of [40] is not listed since it is not publicly available, while others [8], [15], [34], [35], [36], [37], [39], [41] are almost

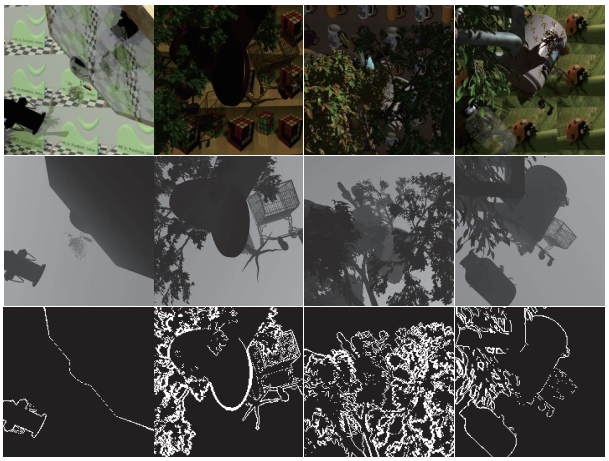
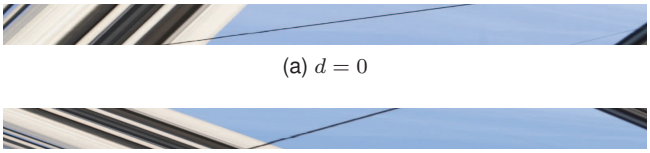


Fig. 11. Light field examples. Top row: central views; Middle row: corresponding depth maps; Bottom row: corresponding occlusion maps.



(a) $d = 0$

(b) $d = 4$

Fig. 12. The sheared EPIs when $d = 0$ and 4, respectively.

real data and do not provide ground truth depth. Fig.14 shows the disparity distributions of the proposed light field in middle resolution. In summary, our dataset has a larger occlusion ratio and disparity range, which is useful for verifying the performance of various EPI based algorithms for the situation that EPI is in "jumping continuity". For each scene, we provide three light fields with different resolutions for further super-resolution tasks. Apart from these, more light fields can be generated with the attached ALFG. These advantages can benefit light field based applications. The new light field dataset and the ALFG will be released on our website [61] upon publication of the work.

We augment the training data using two approaches. The first is exchanging the RGB channels. The second is shearing EPIs [5] using the expression

$$EPI_d(u, x) = EPI_0(u, x + ud), \tag{10}$$

where EPI_0 and EPI_d are the original and sheared EPIs, respectively (Fig.12). The main goal of the shearing operation is to enhance the performance in negative disparity areas.

Note that, the flip operation as commonly used for data augmentation in traditional image super-resolution [51], can not be applied in EPI super-resolution. Note that, as shown in Fig.13a, the intersection between foreground and background is lost after the light field sampling. Thus the flip operation will lead the wrong occlusion to be learnt, causing forbidden occlusion to appear in the reconstructed light field. This is shown in Fig.13c, where an incorrect light field is reconstructed when the foreground is occluded by the background.

(a) The sampled light field (b) Allowed occ. (c) Forbidden occ.

Fig. 13. Flip operation causing the network to learn incorrect occlusions. (a) The sampled light field [62]. The red lines refer to sampled views while the blue/green lines account for the EPI lines; (b) Occlusion after reconstruction; (c) Incorrect occlusion caused by the flip operation.

TABLE 2
Quantitative comparisons between the results yielded by our network with different structures.

Network Structure	4 levels w. LSTM	4 levels w/o. LSTM	3 levels w. LSTM
PSNR(dB)	**28.34**	27.75	28.16
SSIM	**0.886**	0.863	0.884

5.2 Training configuration

The Tensorflow framework [63] is used for training the proposed CNN-LSTM network. The Adam optimizer [64] is chosen to minimize the L_1 loss between the network output and the ground truth. All parameters in the network are initialized using the Xavier initializer [65]. We train the network for 100 epochs. The batch size is 120. The learning ratio is set as $1e-4$ and is reduced by a factor of 0.99 each epoch. The proposed new light field dataset is used to train the network. In total, there are 727200 EPIs used in the training process.

5.3 Synthetic data

To analyze the performance of the proposed network in different disparities, the proposed synthetic light fields are used for validation. In the following experiments, the synthetic light fields with resolution $5\times5\times180\times180$ are super-resolved to $9\times9\times360\times360$. So there is a difference between the maximum disparities in Tab.1 and Fig.14.

5.3.1 LSTM layers

Tab.2 shows the quantitative comparison between the results yielded by our network with and without the LSTM layers. Fig.14 shows a plot of the PSNR for both settings as a function of disparity. Note that, as expected, our network with LSTM layers outperforms the one without LSTM layers over almost all of the disparity range. However, the trend of curves is interesting that it increases with the disparity increases. The main reason for this phenomenon is that the disparities of occluded contents are always smaller than the occluder, so the performance of reconstructed occluded contents is lower than the occluder. As a result, the performance of the reconstruction increases as the disparity increases.

Furthermore, Fig.15 shows qualitative results. Notice that the network with LSTM delivers more details than the one without LSTM in the reconstructed views. This is mainly due to the fact that discrete EPI lines in these areas are over-smoothed as shown in the bottom EPI comparison in Fig.15 when the LSTM is not included. This is consistent with the notion that the LSTM can better cope with the "discontinuous continuity" in EPI.

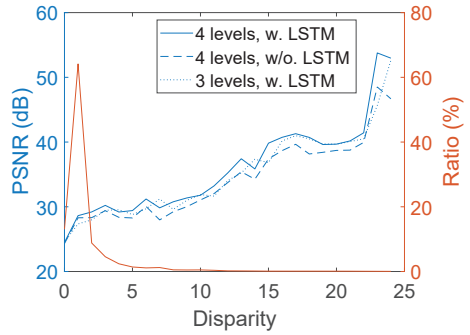


Fig. 14. The histogram of disparity range (reconstructed light fields with 360*360 pixels) and the performance of our network with different structures as a function of disparity.

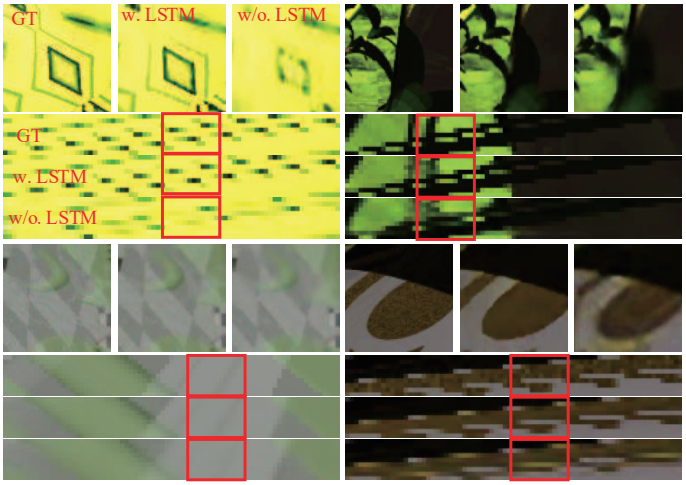


Fig. 15. Qualitative comparisons between the ground truth and the results from networks with and without LSTM layers, respectively. Compared with the one without LSTM, LSTM provides more clear novel views and more accurate EPI lines.

5.3.2 Number of Levels

We remove the 4-th level in Fig.8 to verify the influence of the number of levels. Tab.2 shows the mean PSNR and SSIM over all views. Note that the network with three levels performs weaker than the default one. However, it also outperforms the one without LSTM. In Fig.14, it is noticed that more levels benefit the reconstruction in small disparity areas, so that the network without LSTM outperforms the one with three levels. However, this effect decreases as the disparity increases and the LSTM layers dominate the performance in large disparity areas.

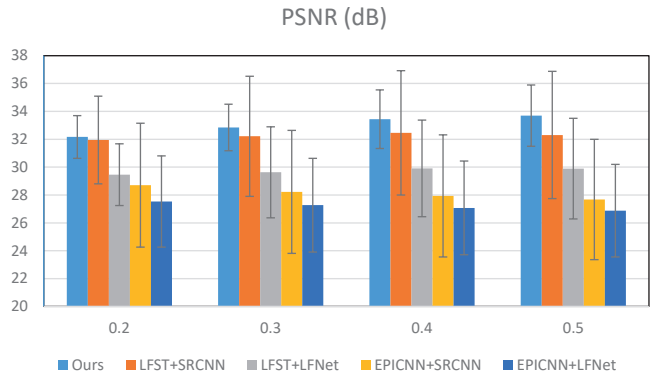
5.4 Real data

5.4.1 Comparison with State-of-the-arts

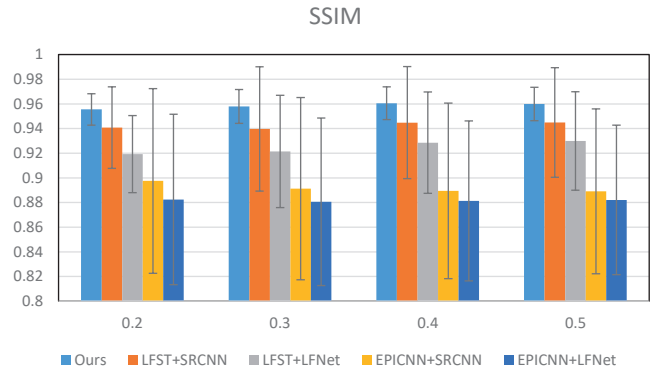
In this experiment, we used the Stanford light field dataset (SLFD) [8]. In order to compare the performance of different methods more fairly, we zoom out each view of the SLFD to $0.2, 0.3, 0.4, 0.5$ of the original size. Recall that the disparity range decreases with respect to the zoom out factor. For the

TABLE 3
Disparity ranges of the SLFD [8] in different zoom out factors.

Factor	0.2	0.3	0.4	0.5
Amethyst	$[-2.4, 2]$	$[-3.6, 3]$	$[-4.8, 4]$	$[-6, 5]$
Bulldozer	$[-1.6, 8]$	$[-2.4, 12]$	$[-3.2, 16]$	$[-4, 20]$
Bunny	$[-3.2, 2]$	$[-4.8, 3]$	$[-6.4, 4]$	$[-8, 5]$
Chess	$[0, 2.8]$	$[0, 4.2]$	$[0, 5.6]$	$[0, 7]$
Lego	$[-3.6, 2.8]$	$[-5.4, 4.2]$	$[-7.2, 5.6]$	$[-9, 7]$
Truck	$[0, 1.2]$	$[0, 1.8]$	$[0, 2.4]$	$[0, 3]$



(a) PSNR (mean and standard deviation)



(b) SSIM (mean and standard deviation)

Fig. 16. Quantitative comparisons with the combinations of the state-of-the-art angular and spatial super-resolution methods. Larger zoom out factor refers to larger disparity range.

reader's reference, the disparity ranges of different sizes are shown in Tab. 3.

Fig.16 shows quantitative comparisons of our method with respect to the alternatives on the SLFD. All values come from all 6 light fields over all views. Note that our method outperforms the alternatives at almost all of the zoom out factors. Although our network only employs synthetic data during the training process, it shows a good generalization ability as applied to unseen camera array data.

SRCNN vs LFNet: It is noticed that the specially designed LFNet for light field spatial super-resolution performs poorly than the single image super-resolution method SRCNN. The main reason is that, the LFNet prefers preserving the angular consistency during the super-resolution process. LFNet will propagate the artifacts in the reconstructed views to other clear input views. Because the SRCNN focuses on super-resolving each view image independently, the spatial super-resolved input views will not

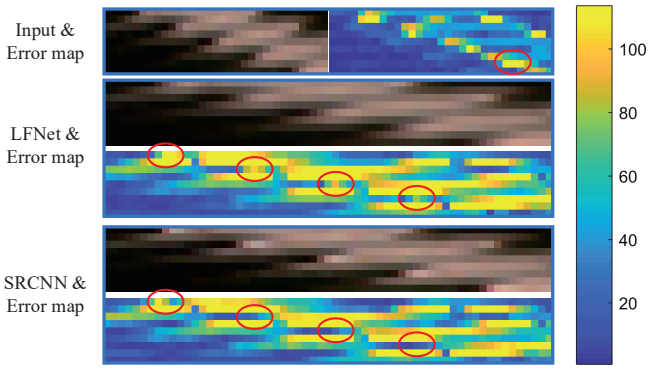


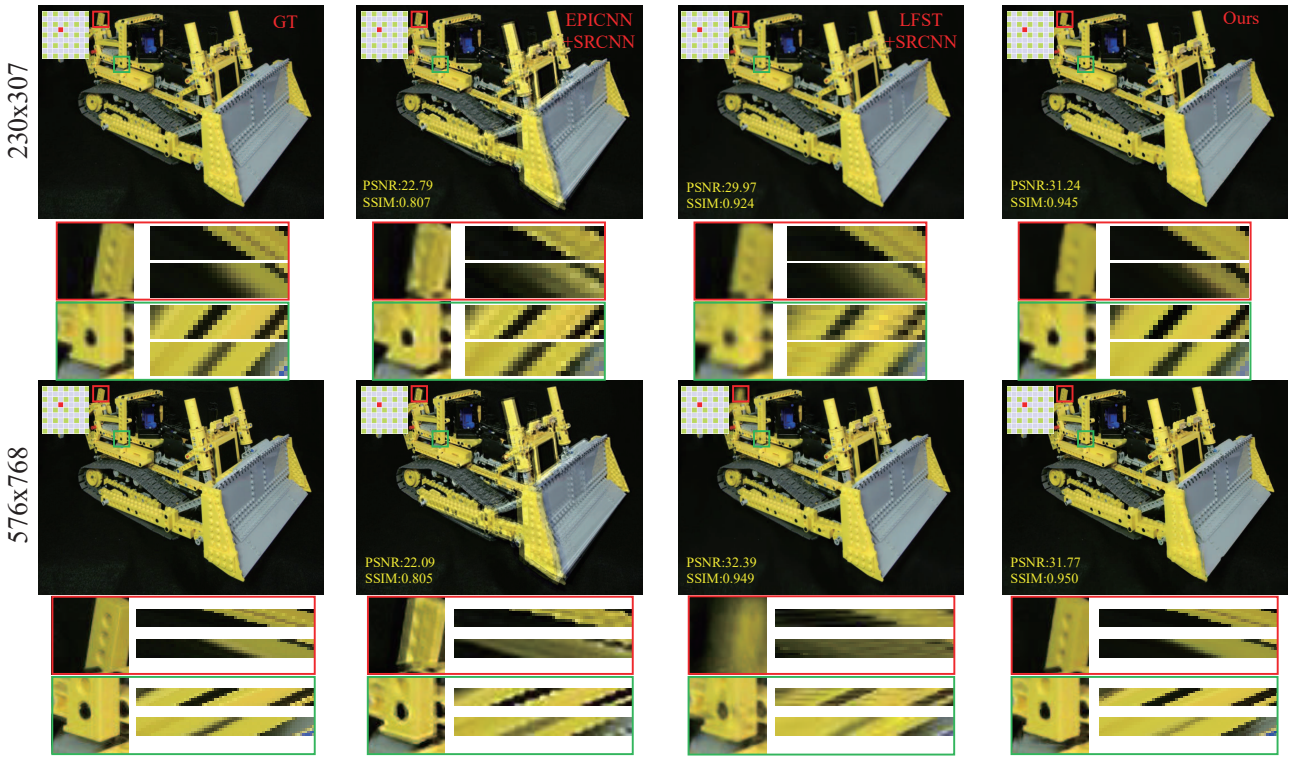
Fig. 17. Qualitative comparisons between the SRCNN and the LFNet with the same input from the LFST. There is a big error in the reconstructed 8-th view from the LFST (the red circle in the the first row). Because the LFNet prefers preserving the angular consistency, this error is propagated to all other views after the super-resolution (red circles in the third row). However, as the SRCNN super-resolves each view image independently, other views are not influenced by the error (red circles in the bottom row).

be influenced by the artifacts in the reconstructed views by the EPICNN or the LFST (see Fig.17). For this reason, we only provide qualitative comparisons with the combination of "EPICNN+SRCNN" and "LFST+SRCNN" later.

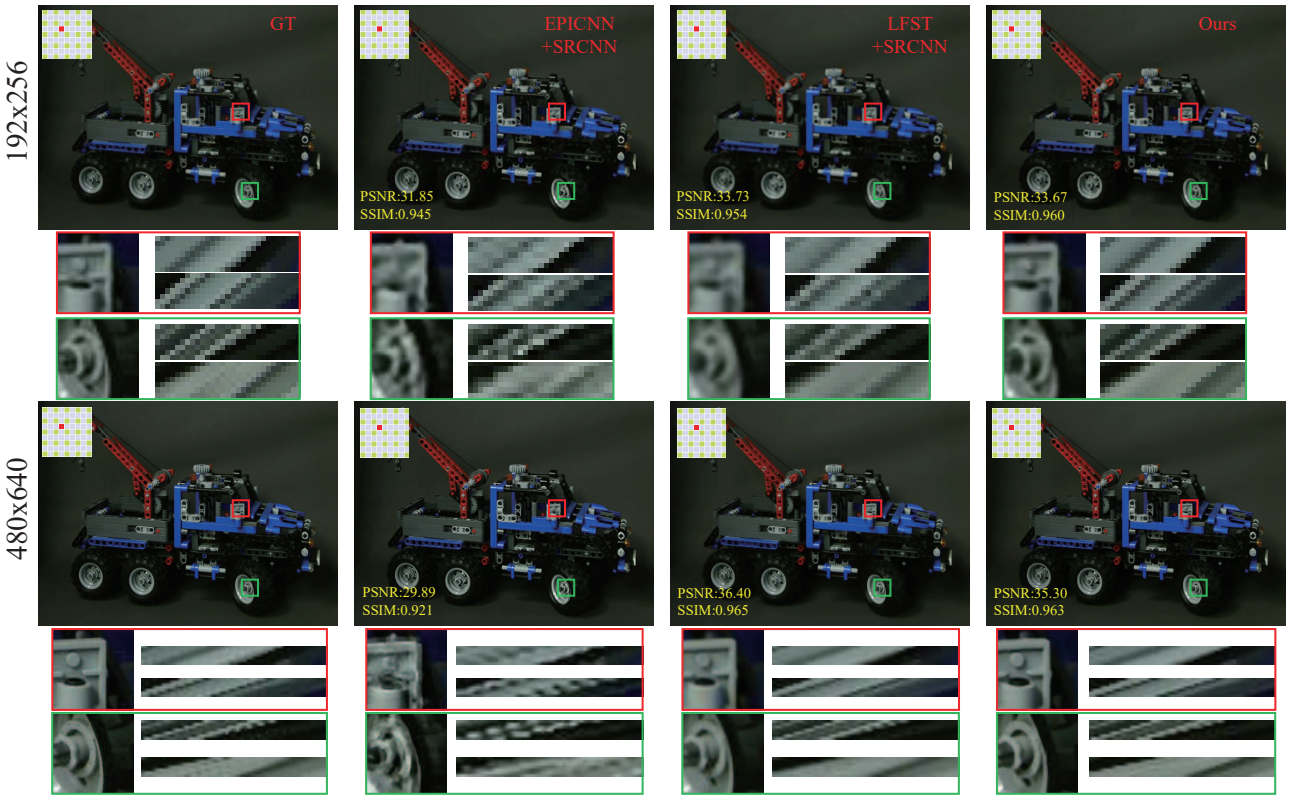
Figs.18,19 show qualitative demonstrations. For each scene in Fig.18, the first and second rows refer to results for zoom out factors 0.2 and 0.5, respectively. Note that EPICNN and LFST achieve in general similar performance as ours at small zoom out factors. However, the performance decreases at large zoom out factors and they tend to over-smooth object boundaries. On the other hand, ours can always maintain sharp object boundaries at both small and large zoom out factors. For example, the boundaries of Bulldozer are all preserved well in Fig.18a. However, previous methods often fail at large zoom out factors. This phenomenon is also well demonstrated in Fig.18b and Fig.19a.

Ours vs EPICNN: Compared with the state-of-the-art methods, our method has achieved at least a 3dB lead (32.17 vs 28.77, in Fig.16a). This advantage increases with the disparity. In the green boxes of Fig.18b, the EPI lines are broken in EPICNN while ours are continuous. Fig.18a gives a better comparison in larger disparity areas. There are serious ghosting in the shovel boundaries recovered by EPICNN. Since the maximum disparity at these areas is about 20 pixels, the EPI consistency on the shovel is lost by EPICNN while our result remains sharp.

Ours vs LFST: Compared with the LFST, the proposed method has two advantages. Firstly, despite LFST achieves good results in large positive disparity regions (such as the shovel boundaries in Fig.18a), the results in negative disparity regions are somewhat mediocre. The best example is Fig.19a, where the areas in front of and behind the toy warriors have positive and negative disparities, respectively. LFST induces ghosting effects in large negative disparity areas. In contrast, our method produces consistent results in both positive and negative disparity areas, thanks to the shearing (Eqn.10) data augmentation and the LSTM's ability to model EPI. Secondly, in contrast with our approach, LFST

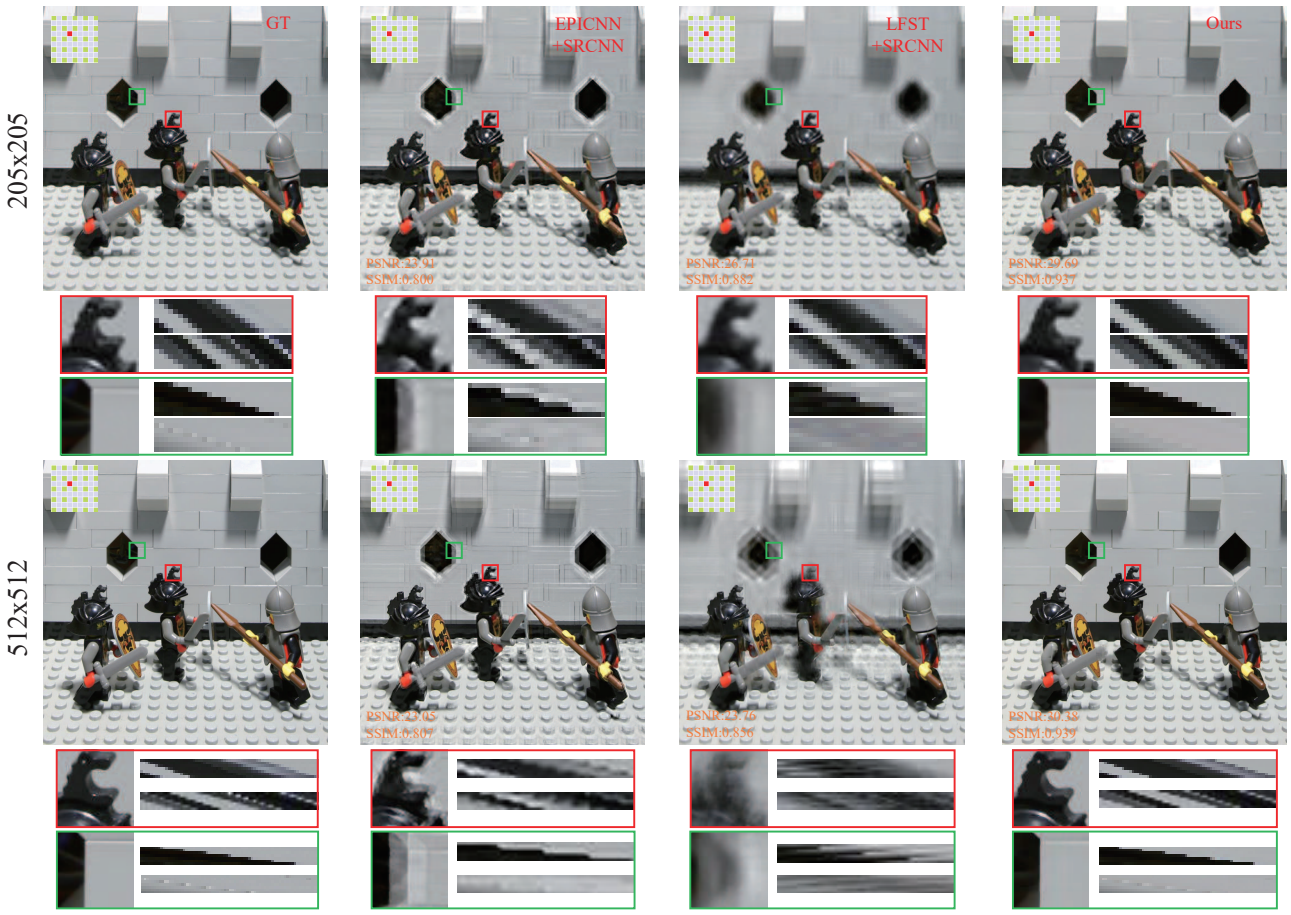


(a) Bulldozer

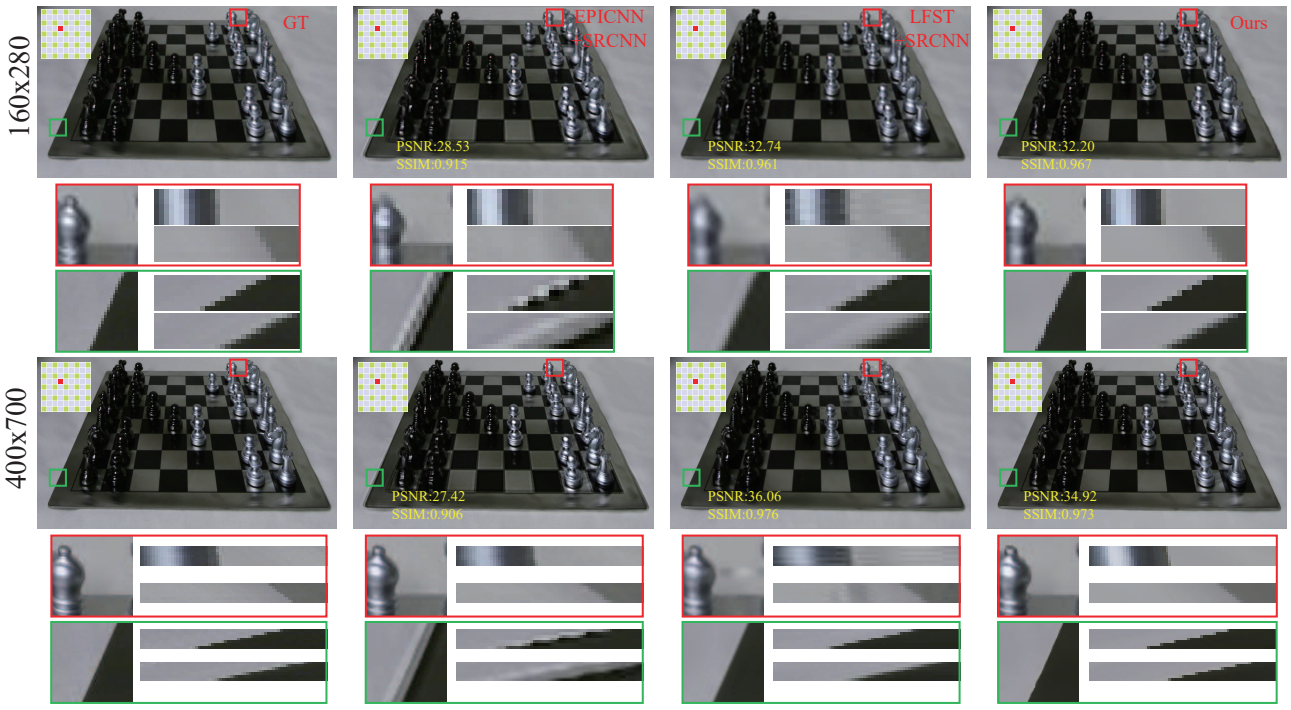


(b) Truck

Fig. 18. Qualitative comparison on the Bulldozer and the Truck. For each light field, the first and second rows show the results at different resolution inputs. For each of the zoom in areas in red and green, the left panel shows the reconstructed view, while the right two rectangular panels show the horizontal and vertical EPIs computed from the reconstructed light field.



(a) Lego



(b) Chess

Fig. 19. Qualitative comparison on the Lego and the Chess. Idem as Fig.18

TABLE 4
Comparisons of our method with the alternatives with fixed spatial resolution and increasing disparity.

Skipped views	Bikes		Couch	
	PSNR(dB)	SSIM	PSNR	SSIM
1 (low dis.)	35.35	0.974	35.59	0.948
3	32.06	0.947	34.23	0.931
5	29.28	0.909	31.52	0.904
7 (high dis.)	28.16	0.893	29.69	0.898

TABLE 5
Comparisons of our method with the alternatives with fixed disparity and increasing spatial resolution.

Skipped views	Bikes		Couch	
	PSNR(dB)	SSIM	PSNR	SSIM
7 (low res.)	29.68	0.923	33.06	0.872
5	30.87	0.942	33.63	0.886
3	32.87	0.963	34.17	0.906
1 (high res.)	35.34	0.975	35.57	0.948

often generates unexpected artifacts in texture boundaries, as shown in some of the green boxes in Fig. 18a and red boxes in Fig.19b[1].

5.4.2 Disparity vs Resolution

Revisiting Fig. 16, it is noted that the performance of the proposed method increases as the disparity increases. To better explain this phenomenon, we conducted another two experiments by fixing disparity range and spatial resolution, respectively. The Disney light field dataset [35] is used here since it has high angular and spatial resolutions. The light fields "Bikes" and "Couch" are selected considering no motion and no other objects in the ground truth.

Disparity: In the first experiment, we fixed the spatial resolution and controlled the disparity range by changing the number of skipped views. Tab.4 shows quantitative comparisons of the proposed method on different disparity ranges. The performance decreases with the increase of disparity. In Fig.20, our method can reconstruct clear EPI structures when the skipped number of views is small (1 and 3, disparities are about 4 and 8.). Because the disparity increases rapidly with the increase of the number of skipped views and there are repeated patterns, the EPI consistency becomes weaker and weaker. It is more and more difficult for our method to reconstruct clear EPI, thus the thin spokes become blur when the disparity becomes large.

Resolution: In the second experiment, we fixed the disparity range and changed the spatial resolution[2]. Tab.5 shows quantitative results on different spatial resolutions with a fixed disparity range. It can be found that a larger spatial resolution leads to a better performance, which is consistent with the tendency of our method shown in Fig. 16. The main reason for this phenomenon is that our method works on EPI domain, *i.e.*, only one spatial dimension is considered. When zooming out the original image to a low-resolution, the distribution of texture is lost in all views. It is difficult to recover a true texture distribution from an incomplete

1. Please watch the provided videos for better comparison.
2. Please refer the supplementary material for more details.

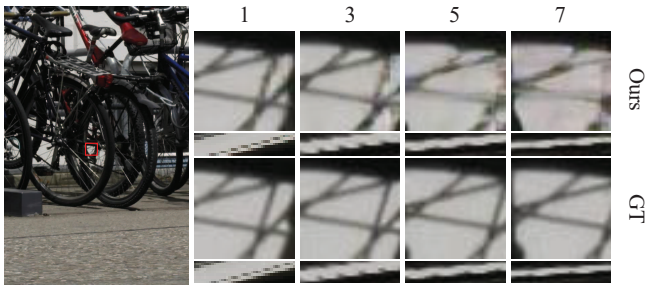


Fig. 20. Results yielded by our method with fixed resolution for several disparity scales.

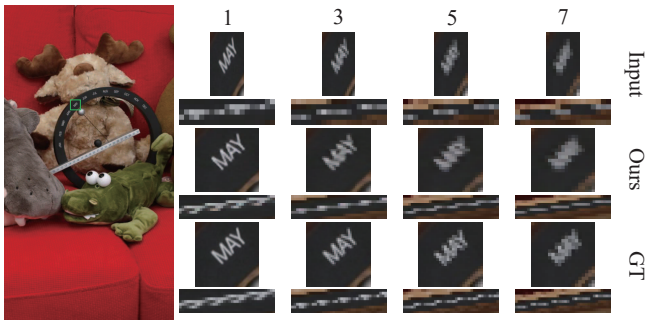


Fig. 21. Results yielded by our method with fixed disparity for several resolution scales.

distribution. Fig.21 demonstrates this phenomenon well. In the lowest resolution version (the rightmost column), it can be found that the text area is observed as continuous gray in all views so that our method produces continuous gray textures in the reconstructed high-resolution light field. While there is a black dot in the ground truth, which damages the distribution of continuous gray color. This problem is not too noticeable in high-resolution light field. Therefore, our method performs better in high-resolution light field compared with the lower one.

In summary, the defects induced in the angular domain can be compensated by the super-resolution in the spatial domain. We guess that the main reason may be the consistency between the spatial and angular patches as analyzed in the [42], [43]. The angular patch is a projection of the spatial patch, thus the proposed simultaneous super-resolution method outperforms the combination of the state-of-the-art light field angular and spatial super-resolution methods. We will explore this issue in the future.

5.5 Discussion

5.5.1 Cross Validation

It is noticed that the performance of our method increases a lot on real data compared with the synthetic one from Tab.2 and Fig.16. We conducted another two experiments to validate the generalization ability of the proposed network.

Training on the SLFD and testing on OurLFs.

We first re-train the proposed network using the SLFD and testing on our synthetic light fields. Tab.6 shows the comparisons. It is noticed that the performance decreases 4dB compared with the results by training on our synthetic light fields. The main reason for this phenomenon is that our

TABLE 6
Quantitative comparisons on OurLFs by training the proposed network on OurLFs and SLFD respectively.

	PSNR(dB)	SSIM
Training on OurLFs	28.34	0.886
Training on the SLFD	24.61	0.789

TABLE 7
Quantitative results by training the proposed network on the mixture LFs and testing on the SLFD.

Zoom out factor	Trainng on OurLFs		Training on the mixture LFs	
	PSNR(dB)	SSIM	PSNR	SSIM
0.2	33.02	0.963	37.02	0.967
0.3	33.78	0.966	38.26	0.970
0.4	34.71	0.968	38.94	0.973
0.5	35.01	0.968	39.08	0.972

light field dataset contains much more challenging occlusion and various lights than the SLFD.

Training on the mixture LFs and testing on the SLFD. In the second experiment, the proposed synthetic light fields and the 'Bulldozer' and 'Lego' from the SLFD are mixed to train the network. The main reason for choosing these two light fields is that they contain the maximum disparity range from Tab.3. Tab.7 shows the results on the other 4 light fields in the SLFD, *i.e.*, the 'Amethyst', 'Bunny', 'Chess' and 'Truck'. It is noticed that the performance increases 4dB, which demonstrates the good generalization ability of the proposed CNN-LSTM network.

5.5.2 Limitation

It is worth noting in passing that, for EPI-driven light field reconstruction, there exists an inherent ambiguity as related to super-resolution. This is because each EPI only covers one spatial dimension whereas every pixel is related to two spatial dimensions in the super-resolution process (*i.e.*, 1 pixel to 2×2 pixels). As shown in Fig.22, the proposed method prefers producing EPI consistency compliant results instead of more clear 2D images in the spatial dimensions. In the future we will explore a method directly operating on the full 4D light field instead of 2D EPIs, aiming to ultimately eliminate such limitation.

6 Conclusion

In this paper, we have indicated that the resolution of a conventional LFC is in fact larger than the number of micro-lenses since most 3D points in a scene are generally defocused. This new insight provides a theoretical basis to overcome the barrier of "spatio-angular trade-off". By analyzing the optical path in an LFC, we have identified the "2D predictable series" nature of the 4D light field. This new feature inspires the introduction of series processing techniques for light field analysis. We have proposed a novel CNN-LSTM network to practically super-resolve a high-resolution light field in both spatial and angular dimensions. Experiments on synthetic and real-world light fields have validated the superiority of the proposed method in large disparity areas.

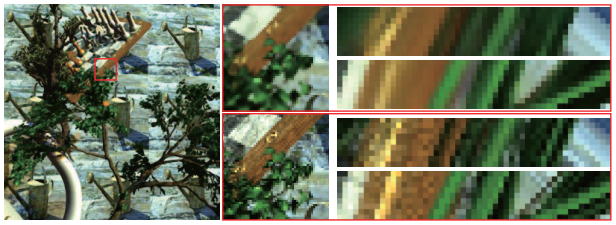


Fig. 22. Limitation of EPI super-resolution.

References

[1] Lytro, "Lytro redefines photography with light field cameras," http://www.lytro.com, 2011.
[2] Raytrix, "∞ raytrix," http://www.raytrix.de, 2012.
[3] M. Levoy and P. Hanrahan, "Light field rendering," in *SIGGRAPH*. ACM, 1996, pp. 31–42.
[4] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *SIGGRAPH*. ACM, 1996, pp. 43–54.
[5] R. Ng *et al.*, *Digital light field photography*. Stanford University, CA, 2006.
[6] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017.
[7] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. K. Nayar, and C. Intwala, "Spatio-angular resolution tradeoffs in integral photography." *Rendering Techniques*, vol. 2006, no. 263-272, p. 21, 2006.
[8] "The new stanford light field archive," http://lightfield.stanford.edu/lfs.html.
[9] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. De Aguiar, N. Ahmed, C. Theobalt, and A. Sellent, "Floating textures," in *CGF*, vol. 27, no. 2. Wiley Online Library, 2008, pp. 409–418.
[10] M. Goesele, J. Ackermann, S. Fuhrmann, C. Haubold, R. Klowsky, D. Steedly, and R. Szeliski, "Ambient point clouds for view interpolation," in *TOG*, vol. 29, no. 4. ACM, 2010, p. 95.
[11] G. Chaurasia, O. Sorkine, and G. Drettakis, "Silhouette-aware warping for image-based rendering," in *CGF*, vol. 30, no. 4. Wiley Online Library, 2011, pp. 1223–1232.
[12] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *TOG*, vol. 32, no. 3, p. 30, 2013.
[13] E. Penner and L. Zhang, "Soft 3d reconstruction for view synthesis," *TOG*, vol. 36, no. 6, p. 235, 2017.
[14] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *TPAMI*, vol. 36, no. 3, pp. 606–619, 2014.
[15] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *TOG*, vol. 35, no. 6, 2016.
[16] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on epi," in *IEEE CVPR*, vol. 2017, 2017, p. 2.
[17] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on epi and extended applications," *IEEE T-PAMI*, vol. 41, no. 7, pp. 1681–1694, 2019.
[18] 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 *IEEE ICCV Workshops*, 2015, pp. 24–32.
[19] T. Georgiev and A. Lumsdaine, "Focused plenoptic camera and rendering," *Journal of Electronic Imaging*, vol. 19, no. 2, p. 021106, 2010.
[20] T. G. Georgiev and A. Lumsdaine, "Superresolution with plenoptic 2.0 cameras," in *Signal recovery and synthesis*. Optical Society of America, 2009, p. STuA6.
[21] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *TPAMI*, vol. 34, no. 5, pp. 972–986, 2012.
[22] M. Broxton, L. Grosenick, S. Yang, N. Cohen, A. Andalman, K. Deisseroth, and M. Levoy, "Wave optics theory and 3-d deconvolution for the light field microscope," *Optics express*, vol. 21, no. 21, pp. 25 418–25 439, 2013.

[23] J. Chang, I. Kauvar, X. Hu, and G. Wetzstein, "Variable aperture light field photography: Overcoming the diffraction-limited spatio-angular resolution tradeoff," in *IEEE CVPR*, 2016, pp. 3737–3745.

[24] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4d rgbd light field from a single image," in *IEEE ICCV*, vol. 2, no. 5, 2017, p. 6.

[25] T.-C. Wang, J.-Y. Zhu, N. K. Kalantari, A. A. Efros, and R. Ramamoorthi, "Light field video capture using a learning-based hybrid imaging system," *TOG*, vol. 36, no. 4, p. 133, 2017.

[26] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *IEEE CVPR*, 2010, pp. 1831–1838.

[27] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous fourier domain," *TOG*, vol. 34, no. 1, p. 12, 2014.

[28] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *TPAMI*, vol. 40, no. 1, pp. 133–147, 2018.

[29] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, "End-to-end view synthesis for light field imaging with pseudo 4dcnn," in *Springer ECCV*, 2018, pp. 340–355.

[30] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modelling of spatial-angular clues," in *Springer ECCV*, 2018, pp. 137–152.

[31] M. Guo, H. Zhu, G. Zhou, and Q. Wang, "Dense light field reconstruction from sparse sampling using residual network," in *Springer ACCV*, 2018, pp. 1–14.

[32] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared EPI structure for light field reconstruction," *TIP*, vol. 28, no. 7, pp. 3261–3273, 2019.

[33] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields." in *VMV*. Citeseer, 2013, pp. 225–226.

[34] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM TOG*, vol. 32, no. 4, p. 46, 2013.

[35] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields." *TOG*, vol. 32, no. 4, pp. 73–1, 2013.

[36] A. S. Raj, M. Lowney, and R. Shah, "Light-field database creation and depth estimation," http://lightfields.stanford.edu/LF2016.html.

[37] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, no. EPFL-CONF-218363, 2016.

[38] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Springer ACCV*, 2016, pp. 19–34.

[39] N. Sabater, G. Boisson, B. Vandame, P. Kerbiriou, F. Babon, M. Hog, R. Gendrot, T. Langlois, O. Bureller, A. Schubert *et al.*, "Dataset and pipeline for multi-view lightfield video," in *IEEE CVPR Workshops*, 2017, pp. 1743–1753.

[40] S. Heber, W. Yu, and T. Pock, "Neural epi-volume networks for shape from light field," in *IEEE CVPR*, 2017, pp. 2252–2260.

[41] R. A. Farrugia and C. Guillemot, "Light field super-resolution using a low-rank prior and deep convolutional neural networks," *arXiv preprint arXiv:1801.04314*, 2018.

[42] T. C. Wang, A. A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE T-PAMI*, vol. 38, no. 11, pp. 2170–2181, 2016.

[43] H. Zhu, Q. Wang, and J. Yu, "Occlusion-model guided anti-occlusion depth estimation in light field," *IEEE J-STSP*, vol. 11, no. 7, pp. 965–978, Oct. 2017.

[44] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *NIPS*, 1990, pp. 396–404.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015, pp. 3431–3440.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[49] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015, pp. 802–810.

[50] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *arXiv preprint arXiv:1805.09817*, 2018.

[51] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *TPAMI*, vol. 38, no. 2, pp. 295–307, 2016.

[52] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *IEEE CVPR*, 2016, pp. 1646–1654.

[53] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE CVPR*, 2017, pp. 624–632.

[54] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *SIGGRAPH*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 307–318.

[55] Z. Lin and H.-Y. Shum, "A geometric analysis of light field rendering," *IJCV*, vol. 58, no. 2, pp. 121–138, 2004.

[56] S. Heber, W. Yu, and T. Pock, "U-shaped networks for shape from light field." in *BMVC*, vol. 3, 2016, p. 5.

[57] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[58] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *TIP*, 2018.

[59] POV-ray, http://www.povray.org/.

[60] G. Tran, "Oyonale - 3d art and graphic experiments," http://www.oyonale.com/.

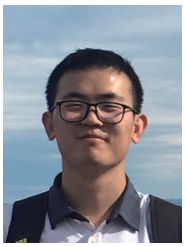
[61] CVPG@NWPU, "Computer vision and computational photography group," http://www.npu-cvpg.org/opensource, 2019.

[62] A. Levin, W. T. Freeman, and F. Durand, "Understanding camera trade-offs through a bayesian analysis of light field projections," in *Springer ECCV*, 2008, pp. 88–101.

[63] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[65] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.



Hao Zhu received B.E. degree from School of Computer Science, Northwestern Polytechnical University in 2014. He is now a Ph.D candidate at School of Computer Science, Northwestern Polytechnical University. His research interests including computational photography, light field computing theory and application.

Mantang Guo received B.E. degree in computer science from school of Computer Science, Northwestern Polytechnical University in 2016. He is now a M.S student at school of Computer Science, Northwestern Polytechnical University. His research interests include computational photography, light field reconstruction and deep learning.

Dr. Hongdong Li is currently a Reader with the Computer Vision Group of ANU (Australian National University). He is also a Chief Investigator for the Australia ARC Centre of Excellence for Robotic Vision (ACRV). He graduated from Zhejiang University with PhD degree, and started as a postdoctoral fellow with the ANU from 2004. His research interests include 3D vision reconstruction, structure from motion, multi-view geometry, as well as applications of optimization methods in computer vision. Prior to 2010, he was with NICTA Canberra Labs working on the "Australia Bionic Eyes" project. He is an Associate Editor for IEEE T-PAMI, and served as Area Chair in recent year ICCV, ECCV and CVPR. He was a Program Chair for ACRA 2015 – Australia Conference on Robotics and Automation, and a Program Co-Chair for ACCV 2018 – Asian Conference on Computer Vision. He won a number of best paper awards in computer vision and pattern recognition, and was the receipt for the CVPR Best Paper Award in 2012 and the Marr Prize Honorable Mention in 2017.

Prof. Qing Wang is now a Professor in the School of Computer Science, Northwestern Polytechnical University. He graduated from the Department of Mathematics, Peking University, in 1991. He then joined Northwestern Polytechnical University. In 1997 and 2000 he obtained Master and PhD degrees in the Department of Computer Science and Engineering, Northwestern Polytechnical University. In 2006, he was awarded as outstanding talent program of new century by Ministry of Education, China. He is now a Senior Member of IEEE and a Member of ACM. He is also a Senior Member of China Computer Federation (CCF). He worked as research scientist in the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University from 1999 to 2002. He also worked as a visiting scholar in the School of Information Engineering, The University of Sydney, Australia, in 2003 and 2004. In 2009 and 2012, he visited Human Computer Interaction Institute, Carnegie Mellon University, for six months and Department of Computer Science, University of Delaware, for one month. Prof. Wang's research interests include computer vision and computational photography, such as 3D reconstruction, object detection, tracking and recognition, light field imaging and processing. He has published more than 100 papers in the international journals and conferences.

Antonio Robles-Kelly received a B.Eng. degree in Electronics and Telecommunications with honours in 1998 and a PhD in Computer Science from the University of York, UK, in 2003. He remained in York until Dec. 2004 as a Research Associate under the MathFit-EPSRC framework and, in 2005, he moved to Australia and took a research scientist appointment with National ICT Australia (NICTA). In 2006 he became the project leader of the Imaging Spectroscopy team at NICTA and, from 2007 to 2009, he was a Postdoctoral Research Fellow of the Australian Research Council. In 2016, he joined CSIRO where he is a Principal Researcher with Data61 and, in 2018, became a Machine Learning and Aritificial Intelligence Professor at Deaking University, Australia. Dr Robles-Kelly's research has been applied to areas such as biosecurity, forensics, food quality assurance and biometrics and is now being deployed by CSIRO under the trademark of Scyllarus (www.scyllarus.com). He has served as the president of the Australian Pattern Recognition Society (APRS) and is an associate editor of the Pattern Recognition Journal and the IET Computer Vision Journal. He is a Senior Member of the IEEE, the president of the TC2 (Technical Committee on structural and syntactical pattern recognition) of the International Association for Pattern Recognition (IAPR) and an Adjunct Associate Professor at the ANU. He has also been a technical committee member, area and general chair of several mainstream computer vision and pattern recognition conferences.