Signal Processing: Image Communication 97 (2021) 116353

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Learning from EPI-Volume-Stack for Light Field image angular super-resolution

Deyang Liu [a], Qiang Wu [b,1], Yan Huang [b], Xinpeng Huang [c,*], Ping An [c,2]

[a] School of Computer and Information, Anqing Normal University, Anqing 246000, China
[b] Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia
[c] Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Keywords:
Light field image angular super-resolution
EPI-volume-stack
3D convolution
Deep learning

ABSTRACT

Light Field (LF) image angular super-resolution aims to synthesize a high angular resolution LF image from a low angular resolution one, and is drawing increased attention because of its wide applications. In order to reconstruct a high angular resolution LF image, many learning based LF image angular super-resolution methods have been proposed. However, most existing methods are based on LF Epipolar Plane Image or Epipolar Plane Image volume representation, which underuse the LF image structure. The LF view spatial correlation and neighboring LF views angular correlations which can reflect LF image structure are not fully explored, which reduces LF angular super-resolution quality. In order to alleviate this problem, this paper introduces an Epipolar Plane Image Volume Stack (EPI-VS) representation for LF angular super-resolution. The EPI-VS is constituted by arranging all LF views in a raster order, which benefits in exploring LF view spatial correlation and neighboring LF views angular correlations. Based on such representation, we further propose an LF angular super-resolution network. 3D convolutions are applied in the whole super-resolution network to better accommodate the input EPI-VS data and allow information propagation between two spatial and one directional dimensions of EPI-VS data. Extensive experiments on synthetic and real-world LF scenes demonstrate the effectiveness of the proposed network. Moreover, we also illustrate the superiority of our network by applying it in scene depth estimation task.

1. Introduction

Light Field (LF) imaging can capture more three-dimensional (3D) information of our world than traditional imaging system [1]. In order to obtain LF image, many hardware systems, such as multi-camera system [2] and time-sequential system [3], have been proposed, with which the intensity and direction information of light rays from real-world scenes can be recorded. Especially, with the introduction of commercial and industrial plenoptic LF cameras (*e.g.*, Lytro [4] and RayTrix [5]), many new applications have been facilitated, such as digital refocusing [6], scene depth estimation [7,8], 3D reconstruction [9], virtual/augmented reality [10], *etc*. However, since the product of spatial resolution and angular resolution cannot exceed sensor resolution, the plenoptic LF camera brings a trade-off problem between spatial and angular resolution, *i.e.*, one can acquire high resolution views in spatial dimensions but only sparse sampling in angular dimensions or vice versa [11].

In order to alleviate this problem and acquire high angular resolution LF content, many computational methods have been proposed focusing on LF angular super-resolution by using a low angular resolution one. Let $L(x, y, u, v)$ denote a light field image, where $x \times y$ is the spatial resolution and $u \times v$ is the angular resolution. Fig. 1 gives an example of reconstructing a high angular resolution LF with angular resolution 7×7 from a low angular resolution LF image with angular resolution 3×3. Generally, LF angular super-resolution methods can be mainly classified into two categories: non-learning based and learning based methods. The non-learning based methods mostly inherit from some traditional solutions proposed for natural 2D images [11]. This kind of method either needs to estimate the depth information of the input sparsely sampled LF views as auxiliary information, or has to use some specific priors, *i.e.* sparsity, in transformation domain [12–17]. This not only easily introduces artifacts in synthesized views, but also limits its applications in practice.

* Corresponding author.
E-mail addresses: deyang.liu@hotmail.com (D. Liu), qiang.wu@uts.edu.au (Q. Wu), yan.huang-3@student.uts.edu.au (Y. Huang), xinpeng_huang@163.com (X. Huang), anping@shu.edu.cn (P. An).
[1] Senior Member, IEEE.
[2] Member, IEEE.

https://doi.org/10.1016/j.image.2021.116353
Received 7 October 2020; Received in revised form 22 April 2021; Accepted 7 June 2021
Available online 12 June 2021

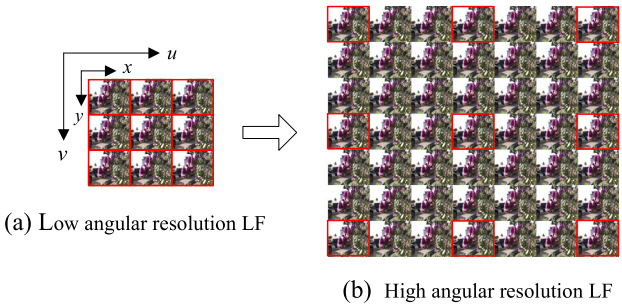

Fig. 1. LF angular super-resolution: (a) Low angular resolution LF image with angular resolution 3 × 3; (b) High angular resolution LF image with angular resolution 7 × 7.

Because of the success of deep leaning, learning based methods [11, 18–25] have achieved promising performance for LF angular super-resolution. However, most of the learning based methods try to reconstruct a high angular resolution LF image based on LF Epipolar Plane Image (EPI) or EPI volume representation. The EPI based LF angular super-resolution methods [11,20,26] try to model the LF angular super-resolution as angular restoration problem in EPI domain. By using a sparsely sampled EPI as input, a high angular resolution EPI can be restored with a convolutional neural network (CNN) based evaluation strategy. The EPI volume based LF angular super-resolution methods intent to synthesize high angular resolution EPI volume in a row or column pattern with sparsely sampled EPI volume as inputs [25,27]. Since LF EPI and EPI volume only contain 2D and 3D information of LF image respectively, the EPI and EPI volume based methods underuse the LF image structure. Moreover, from the extraction process (see Section 2.1), we know that EPI only reflects the changes of the same point in LF views caused by the changes of camera positions in vertical or horizontal direction. The relationships of pixels within each LF view (also called LF view spatial correlation) cannot be fully explored in EPI domain. Meanwhile, since EPI volume only illustrates the relationships of LF views captured by vertical or horizontal directional cameras, the correlations between neighboring LF views in both vertical and horizontal directions (also called LF view angular correlation) also cannot be fully applied by EPI volume based methods. There is still room to improve the LF angular reconstruction quality.

In order to mitigate these problems, we introduce an LF EPI-Volume-Stack (EPI-VS) representation. The EPI-VS is constituted by arranging all LF views in a raster order. Based on the arranging order, the EPI-VS can be further classified into horizontal and vertical EPI-VS. An example of horizontal EPI-VS acquisition is shown in Fig. 2(a). Unlike the EPI or EPI volume representation, EPI-VS is a rearranged version of LF image and can be regarded as a pseudo video representation of LF image data. The EPI-VS has two spatial and one directional dimensions and contains all information of LF image, which can provide benefits in exploring LF view spatial correlation and angular correlation. Moreover, by using a raster order, strong correlations of neighboring LF views in both vertical and horizontal directions can be revealed [28]. Based on LF EPI-VS representation, we further propose an LF angular super-resolution network which is called EPIVS-Net to improve the LF angular super-resolution quality with sparsely-sampled EPI-VS as inputs (see Fig. 2(b)). Since LF view spatial correlation and angular correlation can reflect the LF image inherent structure, they are of great importance for detail texture restoration in LF angular super-resolution. To fully explore such two correlations, the EPIVS-Net adopts 3D convolutional operations in the whole super-resolution procedure. This allows information propagation between EPI-VS two spatial and one directional dimensions. By capturing an interaction between EPI-VS two spatial and one directional dimensions in convolutional operations, the LF view spatial correlation and angular correlation can be fully explored. In addition, the proposed EPIVS-Net can reconstruct high angular resolution LF in one feedforward pass. The main contributions are as follows:

- We introduce a new LF EPI-VS representation for LF image angular super-resolution, which benefits in exploring LF view spatial and angular correlations.
- We propose a learning based EPIVS-Net, in which 3D convolutional operations are adopted in the whole LF super-resolution procedure to better accommodate EPI-VS data and allow information propagation between two spatial and one directional dimensions of EPI-VS data.
- Extensive experiments on real world scenes, synthetic scenes, and LF application validate the effectiveness of the proposed method in achieving a better super-resolution quality and restoring more texture details of synthesized LF views. For some challenging cases, such as occlusions and non-Lambertian surfaces, the proposed method can also outperforms other state-of-the-art approaches.

The rest of this paper is organized as follows. Section 2 reviews the related work of LF angular super-resolution. Problem formulation and low angular resolution EPI-VS construction are described in Section 3. The proposed LF angular super-resolution framework is described in Section 4. Section 5 discusses the simulation results and the last section is devoted to conclusions.

2. Related work

Unlike image spatial super-resolution [29–33], LF image angular super-resolution tries to synthesize high angular resolution LF data with low angular resolution one as input. This section will introduce LF structure and give comprehensive reviews of learning-based LF angular super-resolution methods.

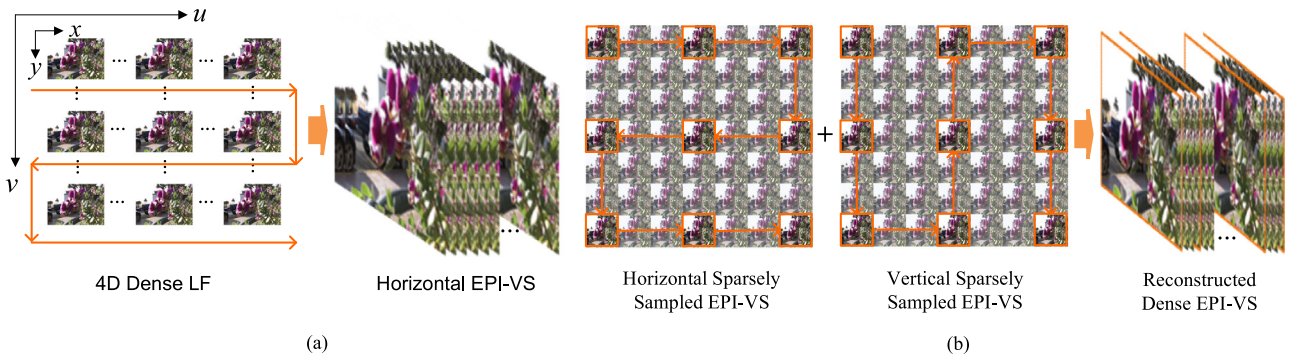

Fig. 2. An illustration of EPI-VS constitution and high angular resolution LF EPI-VS super-resolution: (a) An example of Horizontal EPI-VS constitution, the raster order starts from the first line, horizontally from left to right and then from right to left; (b) High angular resolution LF EPI-VS super-resolution based on horizontal and vertical sparsely-sampled EPI-VS.

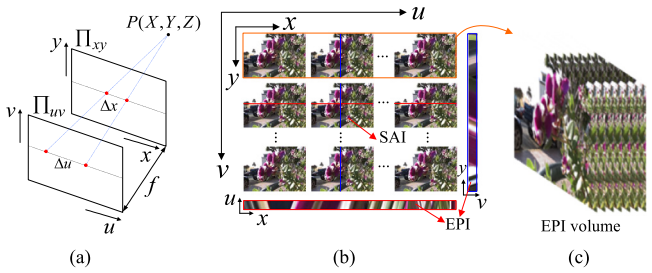

Fig. 3. Illustration of LF structure: (a) Two-plane parameterized model of LF imaging; (b) An example of obtained LF image (SAI array) and extracted EPI; (c) An example of extracted EPI volume from LF data.

2.1. LF structure

LF can be represented by a two-plane parameterized model [34]. Based on the two-plane model, the light ray from any point $P(X, Y, Z)$ in 3D scene travels and intersects at angular position (u, v) in camera plane Π_{uv} and spatial position (x, y) in image plane Π_{xy} to constitute a LF image $L(x, y, u, v)$, as shown in Fig. 3(a). The LF image $L(x, y, u, v)$ can be regarded as a collection of views (also called sub-aperture images (SAIs)) captured by cameras in camera plane with view points parallel to image plane. Since LF image contains four dimensions, LF image is also called 4D LF image in this paper. For a 4D LF image $L(x, y, u, v)$, (x, y) are two spatial dimensions representing two pixel coordinates in each SAI, while (u, v) are two angular dimensions describing camera coordinates in camera plane. Fig. 3 (b) gives an example of obtained LF image. Based on LF structure, we can extract EPI and EPI volume data from LF image. By fixing one angular dimension and one spatial dimension of LF image, *i.e.*, $y = y^*$ and $v = v^*$, an EPI can be extracted (see Fig. 3(b)). IF only one angular dimension is restricted, *i.e.*, $u = u^*$ or $v = v^*$, we can derive an EPI volume $L_{v(u)^*}(x, y, u(v))$ (see Fig. 3(c)). An EPI and an EPI volume only contain 2D and 3D information of LF image, respectively.

2.2. Learning based LF angular super-resolution methods

The success of deep learning in image super-resolution [35–37] has spurred the proposal of many learning-based LF angular super-resolution approaches. Kalantari et al. [18] proposed an end-to-end deep network to synthesize new views with sparse inputs, whose super-resolution process was divided into disparity and color-estimation components. To achieve a better quality of synthesized SAIs, they adopted two sequential CNNs to simultaneously model disparity and color. Vadathya et al. [19] proposed a unified learning framework to reconstruct high angular resolution LF by using minimal number of coded images as inputs. Wu et al. [11] tried to construct a "blur–restoration–deblur" framework to restore the angular details of EPI. However, their method failed in the case of large disparities. This method was subsequently improved by fusing sheared EPIs [20] with a certain value to convert the super-resolution problem to the fusion of sheared EPIs. To solve "either aliasing or blurring" problem in LF angular super-resolution, Wu et al. [26] further proposed a Laplacian Pyramid EPI structure and a corresponding learning based network architecture. Meng et al. [21] proposed to formulate LF super-resolution as a tensor restoration problem and designed a two-stage learning framework to reconstruct high angular resolution LF in spatial and angular dimensions. Yeung et al. [22] proposed an end-to-end LF angular super-resolution network, adopting two subnetworks to characterize the spatial–angular clues in a coarse-to-fine manner using 4D convolutions. However, in their approach, 2D filters were still adopted in the spatial and angular alternating convolution process, which fails to explore the correlation between the spatial and angular dimensions of LF data. By considering the intrinsic geometry information of LF, Jin et al. [23] proposed a learning based network to reconstruct high angular resolution LF with a large baseline, where a depth estimation module and a LF blending module are contained in their model. This paper is further improved by considering sparsely-sampled SAIs with irregular structures [38]. Huang et al. [24] introduced a sequential learning framework to reconstruct high angular resolution LF and estimate depth maps from focal stack images. Wang et al. [27] proposed an end-to-end learning framework making full use of the LF EPI volume representation. With intersection of 2D and 3D convolutions, they constructed a pseudo 4DCNN. This method was subsequently improved by applying a EPI structure preserving loss function in [25]. However, in their methods, only horizontal and vertical SAI correlations are considered, spatio-angular correlation of 4D LF data is still underused.

Since the lost texture details in spatial domain can be recovered in angular domain according to the intrinsic structure of 4D LF [32], it is worth to explore the insight correlations between spatial domain and angular domain to improve the quality of synthesized SAIs, particularly, in occluded regions. However, the EPI and EPI volume based LF super-resolution methods cannot well capture such spatio-angular correlations in LF data. To this end, we propose a new EPI-VS LF representation, which benefits in exploring the spatio-angular correlation of 4D LF data. Moreover, we also put forward a LF angular super-resolution network, where 3D convolutions are adopted in the whole super-resolution procedure to accommodate the EPI-VS data and allow an interaction between LF spatial and angular dimensions.

3. Proposed method

3.1. Problem formulation

A 4D LF $L(x, y, u, v)$ can normally be defined by a two-plane model, where a ray is represented by the interactions of two parallel planes. Suppose that Ω and Π are the two parallel planes. Then the 4D LF can be represented as

$$(x, y, u, v) \rightarrow L(x, y, u, v), \quad L : \Omega \times \Pi \rightarrow \mathbb{R}, \tag{1}$$

where $(x, y)^{\mathrm{T}} \in \Omega$ and $(u, v)^{\mathrm{T}} \in \Pi$ represent spatial and angular coordinates, respectively.

Learning-based LF angular super-resolution can generally be expressed as a task to synthesize a high angular resolution LF $L_{HR}(x, y, u, v)$ from a low angular resolution one $L_{LR}(x, y, u, v)$, which can be written as

$$\begin{aligned} L_{HR}(x, y, u, v) &= f((L_{LR}(x, y, u, v)), \Theta), \\ \Theta^* &= \arg\min_{\Theta} \| L_{GT}(x, y, u, v) - L_{HR}(x, y, u, v) \|_2, \end{aligned} \tag{2}$$

where Θ is the network parameter, $L_{GT}(x, y, u, v)$ is the ground-truth high angular resolution LF image, and $f(\cdot)$ describes the angular mapping from the low angular resolution LF to the high angular resolution LF.

From Eq. (2), we find that 4D convolution may be a straightforward choice to reconstruct dense LF data. However, the high computational burden makes it infeasible for many applications [27]. Therefore, researchers explore to reconstruct dense LF by using LF EPI or EPI volume representation [11,20,25]. We have mentioned above that the EPI and EPI volume are just 2D and 3D slices of the 4D LF, respectively. By using EPI or EPI volume representation, the computational complexity is reduced. However, the LF spatio-angular correlations are not fully explored. To mitigate this problem, we propose an LF EPI-VS representation which is constituted by arranging all the SAIs in a raster order (see Fig. 2(a)). Thus, according to 4D LF $L(x, y, u, v)$, the EPI-VS can be expressed as $\mathscr{L}(x, y, \Psi)$, where Ψ defines the arranging order and numerically equals to $u \times v$. Based on the arranging order, the EPI-VS can be further classified into horizontal EPI-VS $\mathscr{L}(x, y, \Psi_H)$ and vertical EPI-VS $\mathscr{L}(x, y, \Psi_V)$. Compared with the original 4D LF representation, EPI-VS representation reduces the LF dimension from 4D to 3D, which can provide benefits in exploring LF spatio-angular

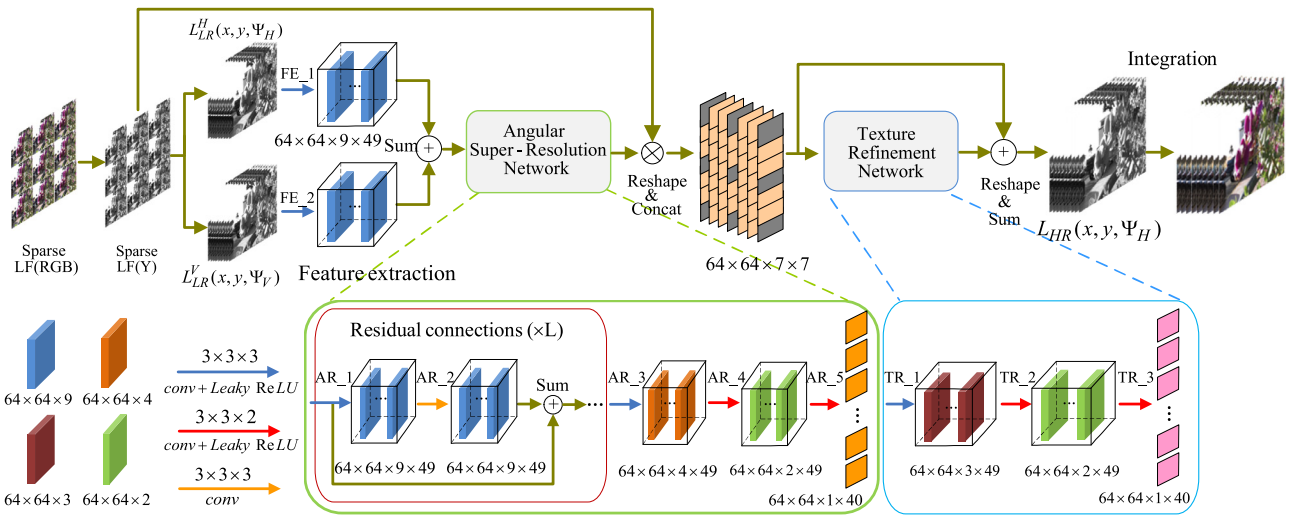

Fig. 4. Overview of proposed network, which aims to reconstruct a high angular resolution LF with angular resolution 7 × 7 from a low angular resolution one with angular resolution 3 × 3. The whole model consists of four parts: feature extraction, angular super-resolution, texture refinement and integration, which allows us to reconstruct a high angular resolution LF in a coarse to fine manner.

correlations. Moreover, since the EPI-VS representation is equivalent to the original 4D LF representation, it is advantageous in keeping the intrinsic LF consistent structure.

Based on the EPI-VS representation, the proposed network focuses on learning a mapping $\mathscr{F}(\cdot)$ to reconstruct high angular resolution EPI-VS (*i.e.,* $\mathscr{L}_{HR}(x,y,\Psi_H)$ or $\mathscr{L}_{HR}(x,y,\Psi_V)$) with horizontal and vertical low angular resolution EPI-VSs (*i.e.,* $\mathscr{L}^H_{LR}(x,y,\Psi_H)$ and $\mathscr{L}^V_{LR}(x,y,\Psi_V)$) as inputs, which can be written as

$$\begin{aligned}&\mathscr{L}_{HR}(x,y,\Psi_{H(V)}) = \mathscr{F}(\mathscr{L}^H_{LR}(x,y,\Psi_H),\mathscr{L}^V_{LR}(x,y,\Psi_V),\boldsymbol{\Phi}),\\&\boldsymbol{\Phi}^* = \arg\min_{\boldsymbol{\Phi}}\|\mathscr{L}_{GT}(x,y,\Psi_{H(V)}) - \mathscr{L}_{HR}(x,y,\Psi_{H(V)})\|_2,\end{aligned} \tag{3}$$

where $\boldsymbol{\Phi}$ is the network parameter, Ψ_H defines the horizontal raster arranging order and Ψ_V defines the vertical raster arranging order, $\mathscr{L}_{GT}(x,y,\Psi_{H(V)})$ is the ground-truth high angular resolution EPI-VS with horizontal or vertical raster arranging order.

After a rearranging of the obtained high angular resolution EPI-VS $\mathscr{L}_{HR}(x,y,\Psi_{H(V)})$, we can derive the high angular resolution LF image. The details of low angular resolution EPI-VS construction and LF angular super-resolution network are given in the following subsections.

3.2. Low angular resolution EPI-VS construction

Suppose that the 4D dense LF data are represented by $L(x,y,u,v)$, as shown in Fig. 1(b). We focus on high angular resolution LF reconstruction by using a low angular resolution LF $L(x,y,u_s,v_s)$ as input, where (u_s,v_s) is the sampled angular position in ground-truth high angular resolution LF. According to the definition of EPI volume, we can obtain horizontal and vertical low angular resolution EPI volumes based on the $L(x,y,u_s,v_s)$. By fixing $v_s = v_s^*$, a horizontal low angular resolution EPI volume $L_{v_s^*}(x,y,u_s)$ is obtained. By analogy, a vertical low angular resolution EPI volume $L_{u_s^*}(x,y,v_s)$ can be derived when $u_s = u_s^*$ is fixed. Since EPI-VS can be regarded as a stack of all EPI volumes with a specific SAI arranging order. Similarly, the low angular resolution EPI-VS is constituted by stacking all low angular resolution EPI volumes with a specific SAI arranging order. For horizontal low angular resolution EPI-VS $\mathscr{L}^H_{LR}(x,y,\Psi_H)$, a horizontal raster arranging order is used, while for vertical low angular resolution EPI-VS $\mathscr{L}^V_{LR}(x,y,\Psi_V)$, a vertical raster arranging order is adopted. An example of low angular resolution EPI-VSs construction in horizontal and vertical directions is shown in Fig. 2(b). Ψ_H and Ψ_V numerically equal to $u_s \times v_s$.

Low angular resolution EPI-VS can provide SAI correlations in two spatial dimensions and one directional dimension, which is useful in improving the LF angular super-resolution quality. Based on the EPI-VS representation, this paper designs a network to reconstruct a high angular resolution LF from a low angular resolution EPI-VS. Given a low angular resolution LF, the horizontal and vertical low angular resolution LF EPI-VSs are firstly obtained. Then the obtained low angular resolution EPI-VSs are fed into the proposed EPIVS-Net to reconstruct the high angular resolution LF EPI-VS.

3.3. LF angular super-resolution network

The proposed EPIVS-Net is illustrated in Fig. 4, which shows an end-to-end mapping from low angular resolution LF to high angular resolution LF. The proposed network focuses on 7 × 7 LF super-resolution with 3 × 3 low angular resolution LF as inputs. In this paper, only 3D convolutional operations are adopted in the proposed network to better accommodate the low angular resolution EPI-VS data. The main advantage is that EPI-VS representation can provide spatio-angular correlations of 4D LF data, and 3D convolution can allow an interaction between the LF spatial and angular dimensions in the LF angular super-resolution process.

As shown in Fig. 4, the low angular resolution SAIs are first converted from RGB space to YCbCr space. In order to reduce the computational complexity, the proposed network is only applied on the luma component (Y), and the other two chrominance components (Cb and Cr) are upsampled using bilinear interpolation. Note that, the proposed EPIVS-Net can also be applied to the Cb and Cr components. In super-resolution process, the horizontal and vertical low angular resolution EPI-VS of sparse LF (Y) are first derived and then fed into the EPIVS-Net after being cropped into small training patches. In this paper, the size of training patches is set to 64 × 64. The proposed EPIVS-Net can be divided into four parts: feature extraction, angular super-resolution, texture refinement, and integration. Feature extraction includes one 3D convolution layer, and is used to extract the 3D features of the input training patches. Here, the 3D features means the high dimensional features containing two spatial and one directional features extracted from input patches. The sizes of input and output tensors of the feature-extraction layer are $64\times64\times d\times c$ and $64\times64\times d\times N$ respectively, where the first three dimensions correspond to two spatial dimensions and one directional dimension of the input training patches respectively. c is the channel dimension, and N is the number of kernels. We adopt a filter size of $3\times3\times3\times1\times49$ in the feature-extraction layer, which means that 49 3D feature maps can be obtained after feature extraction. By executing a 3D convolutional operation, the spatial and angular

Table 1
Detail parameter setting of the proposed network under task $3 \times 3 \rightarrow 7 \times 7$. The first three dimensions of the input and output correspond to two spatial dimensions and one directional dimension. The fourth dimensions of the input and output correspond to channel dimension and number of kernels, respectively.

	Module	Filter size	Input size	Output size	Stride	Padding	Activation
Feature extraction	FE_1	$3 \times 3 \times 3 \times 1 \times 49$	$64 \times 64 \times 9 \times 1$	$64 \times 64 \times 9 \times 49$	[1,1,1]	[1,1,1]	LeakyReLU
	FE_2	$3 \times 3 \times 3 \times 1 \times 49$	$64 \times 64 \times 9 \times 1$	$64 \times 64 \times 9 \times 49$	[1,1,1]	[1,1,1]	LeakyReLU
	Sum	–	$64 \times 64 \times 9 \times 49$ $64 \times 64 \times 9 \times 49$	$64 \times 64 \times 9 \times 49$	–	–	–
Angular super-resolution	AR_1($\times L$)	$3 \times 3 \times 3 \times 49 \times 49$	$64 \times 64 \times 9 \times 49$	$64 \times 64 \times 9 \times 49$	[1,1,1]	[1,1,1]	LeakyReLU
	AR_2($\times L$)	$3 \times 3 \times 3 \times 49 \times 49$	$64 \times 64 \times 9 \times 49$	$64 \times 64 \times 9 \times 49$	[1,1,1]	[1,1,1]	–
	Sum($\times L$)	–	$64 \times 64 \times 9 \times 49$ $64 \times 64 \times 9 \times 49$	$64 \times 64 \times 9 \times 49$	–	–	–
	AR_3	$3 \times 3 \times 3 \times 49 \times 49$	$64 \times 64 \times 9 \times 49$	$64 \times 64 \times 4 \times 49$	[1,1,2]	[1,1,0]	LeakyReLU
	AR_4	$3 \times 3 \times 2 \times 49 \times 49$	$64 \times 64 \times 4 \times 49$	$64 \times 64 \times 2 \times 49$	[1,1,2]	[1,1,0]	LeakyReLU
	AR_5	$3 \times 3 \times 2 \times 49 \times 40$	$64 \times 64 \times 2 \times 49$	$64 \times 64 \times 1 \times 40$	[1,1,2]	[1,1,0]	–
	Reshape & Concat	–	$64 \times 64 \times 9 \times 1$ $64 \times 64 \times 1 \times 40$	$64 \times 64 \times 7 \times 7$	–	–	–
Texture refinement	TR_1	$3 \times 3 \times 3 \times 7 \times 49$	$64 \times 64 \times 7 \times 7$	$64 \times 64 \times 3 \times 49$	[1,1,2]	[1,1,0]	LeakyReLU
	TR_2	$3 \times 3 \times 2 \times 49 \times 49$	$64 \times 64 \times 3 \times 49$	$64 \times 64 \times 2 \times 49$	[1,1,1]	[1,1,0]	LeakyReLU
	TR_3	$3 \times 3 \times 2 \times 49 \times 40$	$64 \times 64 \times 2 \times 49$	$64 \times 64 \times 1 \times 40$	[1,1,1]	[1,1,0]	LeakyReLU
	Reshape & Sum	–	$64 \times 64 \times 7 \times 7$ $64 \times 64 \times 1 \times 40$	$64 \times 64 \times 1 \times 49$	–	–	–
Integration	Y Cb Cr	–	$64 \times 64 \times 1 \times 49$ $64 \times 64 \times 1 \times 49$ $64 \times 64 \times 1 \times 49$	$64 \times 64 \times 3 \times 49$	–	–	–

dimensions of 4D LF achieve an interaction. The obtained 3D feature maps are later added together, and then fed into a angular super-resolution network and texture refinement network to synthesize a high-quality luma component of the high angular resolution EPI-VS. Subsequently, the upsampled chrominance components (Cb and Cr) are integrated with the Y component in the integration part to construct the final high angular resolution EPI-VS data.

Below, we provide details of the angular super-resolution network, texture refinement network, local residual learning method, and training details. The detail parameter setting of the proposed network under task $3 \times 3 \rightarrow 7 \times 7$ (reconstruct 7×7 high angular resolution LF with 3×3 low angular solution one as input) is shown in Table 1.

3.3.1. Angular super-resolution network

The angular super-resolution network aims to synthesize an intermediate novel LF SAIs by performing 3D convolutional operations with interaction in the spatial–angular dimensions. The angular super-resolution network contains one cyclic residual connection layer and three 3D convolution layers. The obtained feature maps are fed into a cyclic residual block (residual connection) layer, as shown in Fig. 4. Each residual block performs one $3 \times 3 \times 3$ 3D convolutions followed by a leaky ReLU with $\alpha = 0.2$ and one 3D convolution without leaky ReLU. The size of 3D convolution filters in each residual block is set to $3 \times 3 \times 3 \times 49 \times 49$, with padding of 1 and stride of 1. After the cyclic residual block, we obtain a feature map of size $64 \times 64 \times 9 \times 49$. Here '9' means the derived feature map has nine channels, and '49' can be regarded as the number of SAIs at the final residual block layer. To derive the intermediate novel SAIs after the cyclic residual connection layer, we downsample the channel dimensions from 9 to 1. Therefore, three more 3D convolutional layers are applied. The first 3D convolutional layer has a kernel with size $3 \times 3 \times 3$, padding of 0 in directional dimension, and stride of 2, while the kernel size of the other two layers is $3 \times 3 \times 2$, padding of 0 in directional dimension, and stride of 2. Thus, the filter sizes of the three layers are $3 \times 3 \times 3 \times 49 \times 49$, $3 \times 3 \times 2 \times 49 \times 49$, and $3 \times 3 \times 2 \times 49 \times 40$, respectively. After the first two 3D convolutional layer, the output has a size of $64 \times 64 \times 2 \times 49$. By performing the third 3D convolutions, we can derive the coarse intermediate novel LF SAIs with size $64 \times 64 \times 1 \times 40$. Subsequently, the derived coarse intermediate novel LF SAIs and input sparse LF SAIs are concatenated together according to the corresponding coordinates in the original dense LF data to construct the intermediate high angular resolution LF of size $64 \times 64 \times 7 \times 7$ with a *Reshape* operation.

3.3.2. Texture refinement network

In the LF angular super-resolution procedure, the high-frequency texture is difficult to restore due to information asymmetry [11]. Therefore, inspired by LF super-resolution methods [21,22], we propose a texture refinement network to further recover more high-frequency details of the reconstructed intermediate novel SAIs. The texture refinement network attempts to learn a transformation $\mathcal{F}(\cdot)$ to recover the high-frequency texture $\mathcal{F}(S)$ for the reconstructed intermediate novel SAIs S, which can be expressed as

$$\hat{S} = S + \mathcal{F}(S), \tag{4}$$

where $\hat{S}$ denotes the final reconstructed novel SAIs.

The detail of refinement network is shown in Fig. 4. The input is organized into $64 \times 64 \times 7 \times 7$ dimensions after concatenation and reshaping operations. To obtain $\mathcal{F}(S)$ of size $64 \times 64 \times 1 \times 40$, three 3D convolutional layers are adopted to reduce the angular dimension (the channel dimension) by a factor of two. The kernel size of the first 3D convolutional layer is $3 \times 3 \times 3$ and the other two layers have kernels with size $3 \times 3 \times 2$. The first two layers perform a 3D convolution, followed by a leaky ReLU with $\alpha = 0.2$, while the third layer only performs a 3D convolution without using activation function. The filter sizes of the three layers are $3 \times 3 \times 3 \times 7 \times 49$, $3 \times 3 \times 2 \times 49 \times 49$, and $3 \times 3 \times 2 \times 49 \times 40$, respectively, with padding of 0 in directional dimension and stride of 2. The estimated $\mathcal{F}(S)$ is then added to the intermediate dense LF to constitute the final high angular resolution EPI-VS.

4. Experimental setup

4.1. Training details

The proposed network is designed to synthesize desired dense 4D LF data with minimum perceptual loss. To this end, we adopt a loss function by directly minimizing the $\mathcal{L}_2$ distance between the final synthesized EPI-VS $\mathscr{L}_{HR}(x, y, \Psi_{H(V)})$ and the corresponding ground-truth $\mathscr{L}_{GT}(x, y, \Psi_{H(V)})$, which can be expressed as

$$\mathcal{L}_2 = \sum_x \sum_y \sum_{\Psi} (\mathscr{L}_{HR}(x, y, \Psi_{H(V)}) - \mathscr{L}_{GT}(x, y, \Psi_{H(V)}))^2. \tag{5}$$

In the training process, we select 75 LF scenes from publicly available datasets [18] for training. All the LF scenes are captured by a Lytro Illum camera, and decoded in 14×14 SAIs with resolution

Table 2
Training parametric statistics of proposed EPIVS-Net.

Training memory usage	Training time	Model size
1376 MB	3–4 Days	9.5 MB

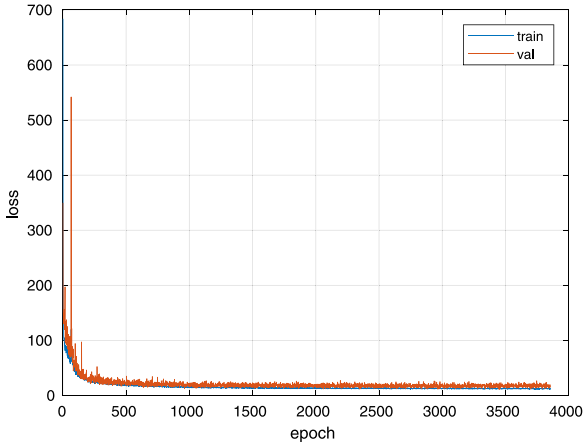

Fig. 5. Curves of training loss and validation loss against the number of epoch.

376×541. Considering the impact of vignetting and optical distortion, only the central 7×7 SAIs are selected in the experiment. Patches with spatial resolution 64×64 are extracted from each SAI with stride 1 to construct the training dataset and the number of total training samples is over 11,000,000. The end-to-end model is optimized by the stochastic gradient descent method with batch size of 1.

The proposed model is implemented using MatConvNet toolbox [39]. A Gaussian distribution [40] with standard deviation $\sqrt{2/N}$ is used to initialize the weights of the network, and all the biases are initialized to zero. The learning rate is set to $1e-6$, and decreased by a factor of 2 for every 1000 epochs. Each epoch contains 1000 iterations. Fig. 5 gives the curves of training loss and validation loss against the number of epoch under task $2 \times 2 \rightarrow 7 \times 7$, and the network converges nearly 3800 epoches.

Many image super-resolution quality evaluation metrics have been put forward [41]. Since the proposed EPIVS-Net is a strong pixel-level supervision method, where the pixel-wise $\mathcal{L}_2$ distance is adopted in the proposed loss function, the Peak Signal to Noise Ratio (PSNR)and Structural Similarity Index Measure (SSIM) are adopted as the quality evaluation metrics.

Table 2 gives some training parametric statistics of proposed EPIVS-Net. In order to reduce the memory usage, we extract patches with resolution 64×64 from each SAI with stride 1 to establish the training dataset. Moreover, the batch size is also set to 1 to decrease the memory usage for training phases.

4.2. Local residual learning method

A cyclic residual connection network is adopted in the angular super-resolution network. To stabilize the training procedure, three skip connection methods borrowed from [36] are examined for the residual connection network, as shown in Fig. 6.

- ***No skip connection:*** The residual connection network uses no local skip connection.
- ***Shared-source skip connection:*** The residual connection network adopts the local skip connection proposed by DRRN [42].
- ***Sequential skip connection:*** The ResNet-style local skip connection [43] is used in the residual connection network.

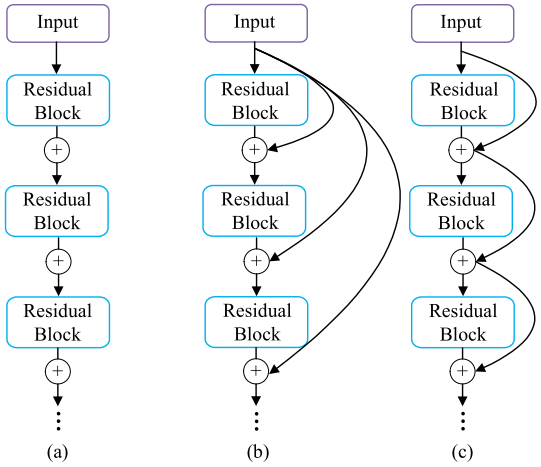

Fig. 6. Illustration of three local residual connections: (a) No skip connection; (b) Shared-source skip connection; (c) Sequential skip connection.

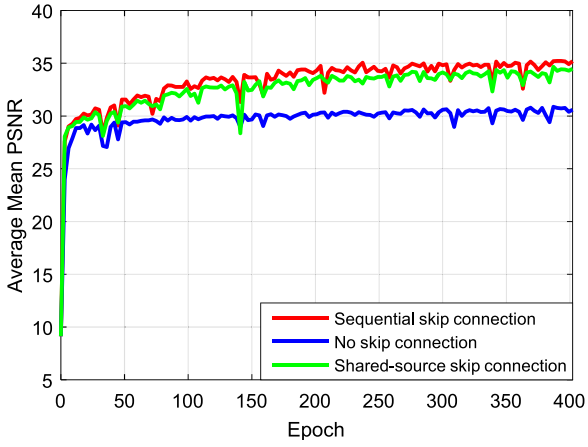

Fig. 7. Convergence analysis on three types of connection. For each connection, the curve is derived on a selected validation set based on average PSNR of reconstructed novel SAIs.

Fig. 7 shows the convergence curves of three types of connection in terms of average PSNR of reconstructed novel SAIs on a selected validation set. We find that the *sequential skip connection* always converges to a better point. Hence we select it as the skip connection method in our residual connection network.

5. Experimental results

Extensive experiments are conducted on synthetic [44] and real-world LF datasets [18,45] to illustrate the effectiveness of the proposed method. The proposed EPIVS-Net is firstly tested on task $3 \times 3 \rightarrow 7 \times 7$, which means reconstruction of an 7×7 high angular resolution LF with 3×3 LF SAIs as input (see Fig. 1). In EPIVS-Net, the number of cyclic residual connection layers is set to 8. Comparisons are made on state-of-the-art learning-based LF angular super-resolution methods, including P4DCNN proposed by Wang et al. [27], SACNet proposed by Yeung et al. [22], EPICNN proposed by Wu et al. [11] and EP4DCNN proposed by Wang et al. [25]. In order to illustrate that the proposed network also works well for large disparity cases, we further compare our network with LF angular resolution methods, including LBVSNet proposed by Kalantari et al. [18], SACNet proposed by Yeung et al. [22], EPICNN proposed by Wu et al. [11], and SEPINet proposed by Wu et al. [20] on task $2 \times 2 \rightarrow 7 \times 7$. For synthetic scenes, comparisons are made with SACNet [22] and EPICNN [11] to illustrate the robustness

Table 3
Quantitative comparisons (PSNR/SSIM) of the EPIVS-Net with the state-of-the-art ones under task $3\times3 \rightarrow 7\times7$.

	P4DCNN [27]	SACNet [22]	EPICNN [11]	EP4DCNN [25]	EPIVS-Net
30 Scenes	43.28/0.9912	43.61/0.9873	41.04/0.9782	43.82/**0.9926**	**44.04**/0.9889
Reflective	38.39/0.9508	43.51/0.9778	41.24/0.9703	39.93/0.9594	**43.74**/**0.9801**
Occlusions	32.14/0.9002	40.01/**0.9613**	36.85/0.9382	34.69/0.9231	**40.23**/0.9566
Average	37.94/0.9474	42.38/**0.9755**	39.71/0.9622	39.48/0.9584	**42.67**/0.9752

of EPIVS-Net. To verify that the proposed method can well restore more texture details, comparisons on depth estimation performance are also provided. The average PSNR and SSIM of reconstructed SAIs against the ground truth are adopted as performance criteria. For LBVSNet [18], SACNet [22] and SEPINet [20], we retrain the released models with the same training data, and comparable performances are achieved for comparisons. As to the P4DCNN [27], EP4DCNN [25] and EPICNN [11], the pre-trained models provided by authors are used for comparisons.

5.1. Real-world scenes

Three LF datasets comprising over 110 real-world LF images, which we call *30 scenes* [18], *Occlusions* [45], and *Reflective* [45], are used to evaluate EPIVS-Net. For SACNet [22], the number of spatial–angular alternating convolutional layers is set to 8.

Table 3 shows the quantitative comparisons of EPIVS-Net with the state-of-the-art methods under task $3\times3 \rightarrow 7\times7$. From Table 3, we find that EPIVS-Net achieves a higher average PSNR than the other methods. More specifically, EPIVS-Net achieves average increases on PSNR by 4.73 dB, 0.29 dB, 2.96 dB and 3.19 dB, respectively, which outperform P4DCNN [27], SACNet [22], EPICNN [11] and EP4DCNN [25], accordingly. There are two main reasons for this. First, EPICNN [11], P4DCNN [27] and EP4DCNN [25] all try to reconstruct dense LF by exploring LF EPI or EPI volume representation. Since the EPI and EPI volume are only 2D and 3D slices of a 4D LF, the intrinsic consistent structure of the reconstructed dense LF is overlooked. Moreover, the LF SAI spatial geometry correlation and inter-SAI correlation across the angular dimensions are not fully explored, which also reduce the reconstruction quality. Second, SACNet [22] tries to constitute a network by exploring the high-dimensional LF structure. But it fails to capture an interaction between the spatial and angular dimensions in spatial–angular alternating convolution, which cannot well recover the lost texture details in synthesized SAIs. Unlike these methods, the proposed method introduces a LF EPI-VS representation, which benefits in exploring spatio-angular correlations of 4D LF data. By using 3D convolutional operations in whole LF angular super-resolution procedure, we can fully explore SAI spatial geometry correlation and inter-SAI correlation across the LF angular dimensions and achieve a better super-resolution quality.

Fig. 8 shows the visual comparisons of the reconstructed SAI using different methods. These results indicate that EPIVS-Net can derive reconstructed SAI with a better perceptual quality than the other methods in some texture areas, especially for cases of reflection and occlusion regions. For example, *IMG_1528_eslf* contains complicated textures around the market sign and leaves areas, which are challenging for reconstruction. From the close-up images and error maps, we find the proposed method can reconstruct more texture details than the other methods. The *Occlusions_30_eslf* contains complicated occlusions, which are difficult to reconstruct. Some artifacts are introduced by the compared methods. By contrast, our method obtains a better reconstruction quality. Regarding *Reflective_30_eslf*, EPIVS-Net also achieves a better visual quality than the other methods in reflective surface areas, as found from close-up images and the error maps.

Table 4
Quantitative comparisons (PSNR/SSIM) of EPIVS-Net with Yeung *et al.* [22] and Wu *et al.* [11] on synthetic LF scenes under task $3\times3 \rightarrow 7\times7$.

	SACNet [22]	EPICNN [11]	EPIVS-Net
Sideboard	31.08/0.9165	29.91/0.8951	**31.92**/**0.9321**
Kitchen	35.77/0.9207	34.45/0.8992	**36.53**/**0.9345**
Museum	35.40/0.9106	34.83/0.8983	**36.03**/**0.9220**
Pens	34.88/0.8599	34.02/0.8271	**35.73**/**0.8888**
Platonic	39.14/0.9549	37.56/0.9323	**40.60**/**0.9699**
Tomb	39.66/0.9225	38.96/0.9070	**40.40**/**0.9350**
Town	40.00/0.9738	37.09/0.9567	**40.32**/**0.9764**
Vinyl	39.58/0.9751	37.28/0.9632	**40.51**/**0.9802**
Average	36.94/0.9293	35.51/0.9096	**37.76**/**0.9424**

5.2. Synthetic scenes

To verify that the proposed network can be applied on synthetic LF scenes, we adopted eight synthetic LF scenes from the HCI datasets [44] for comparison. The selected scenes contain different texture features, in which *Kitchen*, *Museum* and *Vinyl* contain some reflective regions, *Platonic, Tomb*, and *Town* contain more complex textures, and *Pens* and *Sideboard* contains some occlusions.

Table 4 shows quantitative comparisons of EPIVS-Net, SACNet [22] and EPICNN [11] on synthetic LF scenes under task $3\times3 \rightarrow 7\times7$. We observe that EPIVS-Net is superior to the other compared methods on all the test LF scenes. Compared to SACNet [22] and EPICNN [11], the average PSNR gains are up to 0.82 dB and 2.25 dB, respectively. This means that the proposed method works better than the two compared methods on large baseline synthetic LF scenes. For *Kitchen*, *Museum* and *Vinyl*, the average PSNR gains over SACNet [22] and EPICNN [11] are 0.77 dB and 2.17 dB, which illustrate that the proposed EPIVS-Net is superior to the other two methods on large baseline synthetic LF scenes with reflective regions. For *Pens* and *Sideboard*, average PSNR gains over SACNet [22] and EPICNN [11] are 0.85 dB and 1.86 dB, which verifies the effectiveness of EPIVS-Net on large baseline synthetic LF scenes with occlusions. For the high texture LF scenes, *i.e.*, *Platonic, Tomb* and *Town*, the average PSNR gains over SACNet [22] and EPICNN [11] are 0.84 dB and 2.57 dB, which implies that EPIVS-Net can reconstruct a high-quality dense LF for large baseline synthetic LF scenes with high textures.

Fig. 9 shows visual comparisons of reconstructed SAI for synthetic scenes using EPIVS-Net, SACNet [22] and EPICNN [11] under the task $3\times3 \rightarrow 7\times7$. From Fig. 9, we see that the proposed EPIVS-Net can achieve a better reconstruction quality. For example, the results of SACNet [22] and EPICNN [11] show some blurring and ghosting artifacts for *Kitchen*, *Sideboard* and *Vinyl*. Especially for some texture areas, such as character areas of basketball and books in *Sideboard* and *Vinyl*, the reconstructed SAIs are quite blurry. By contrast, the proposed EPIVS-Net can restore more texture details and achieve a higher reconstruction quality, which can be seen from the close-up images and error maps.

5.3. Comparisons on task $2\times2 \rightarrow 7\times7$

To further illustrate the robustness of EPIVS-Net on real-world LF scenes with large disparity, we compare it with four state-of-the-art methods, including LBVSNet [18], SACNet [22], EPICNN [11], and SEPINet [20], under the task $2\times2 \rightarrow 7\times7$.

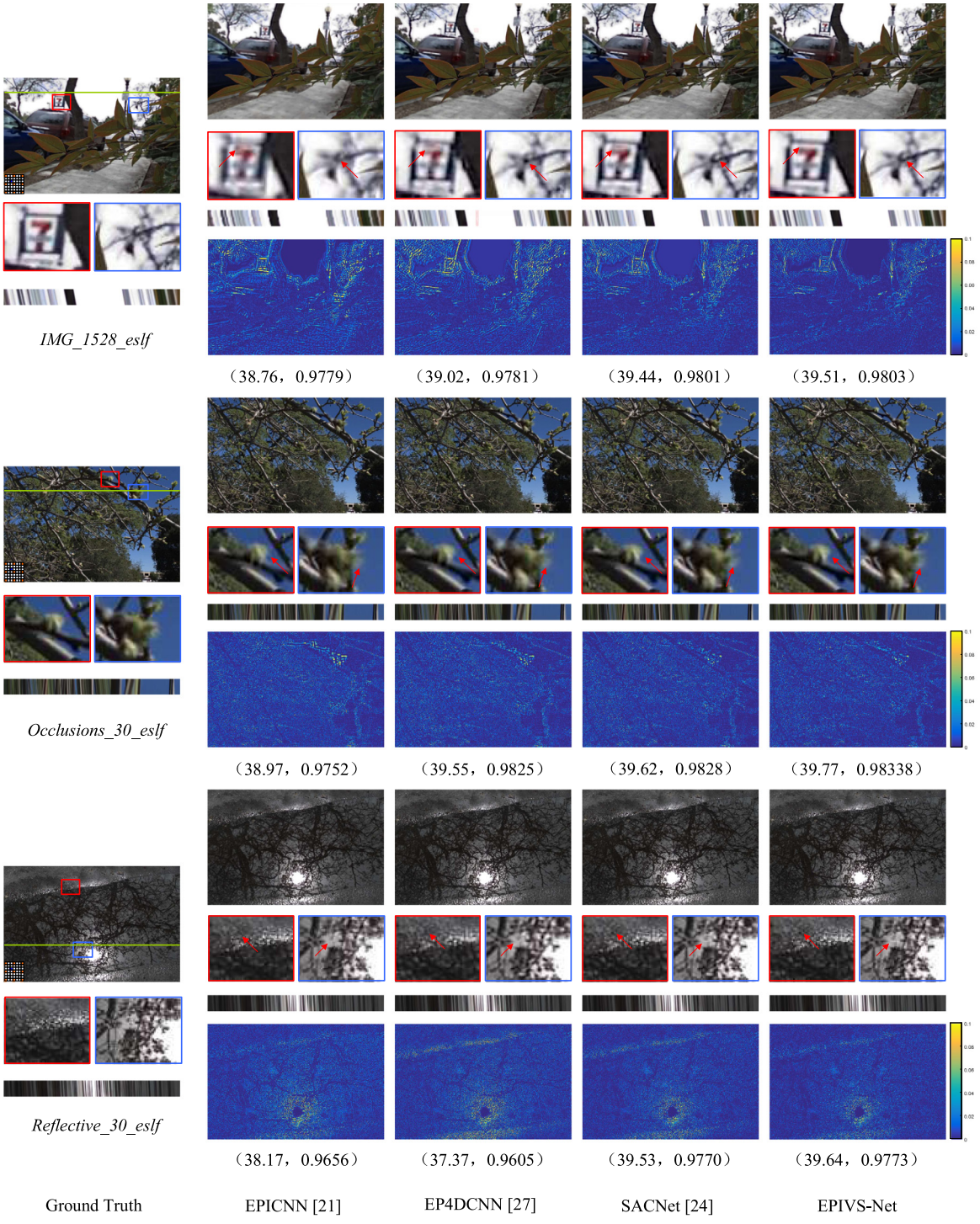

Fig. 8. Visual comparisons of reconstructed SAI using different methods for the task $3 \times 3 \rightarrow 7 \times 7$. The comparison shows ground-truth SAIs, error maps of reconstructed SAIs in Y channel, close-up versions of SAI portions in blue and red boxes, and extracted EPIs at the green line.

Table 5
Quantitative comparisons (PSNR/SSIM) of EPIVS-Net 8L with three state-of-the-art methods under task $2 \times 2 \rightarrow 7 \times 7$. Input 2×2 sparse SAIs are sampled at four corners.

	LBVSNet [18]	SACNet [22]	EPICNN [11]	SEPICNN [20]	EPIVS-Net
30 Scenes	38.83/0.976	41.19/0.981	33.66/0.918	39.17/0.975	**42.14/0.985**
Reflective	36.44/0.946	39.90/0.966	32.72/0.924	36.38/0.944	**40.49/0.969**
Occlusions	32.92/0.922	36.15/0.942	34.76/0.930	34.41/**0.955**	**36.90**/0.945
Average	36.06/0.948	39.08/0.963	33.71/0.924	36.65/0.958	**39.84/0.966**

Table 5 shows quantitative comparisons of EPIVS-Net with other four state-of-the-art methods under task $2 \times 2 \rightarrow 7 \times 7$. From Table 5, we see that EPIVS-Net always has higher average PSNR and SSIM values than the other methods. EPIVS-Net has average increases of

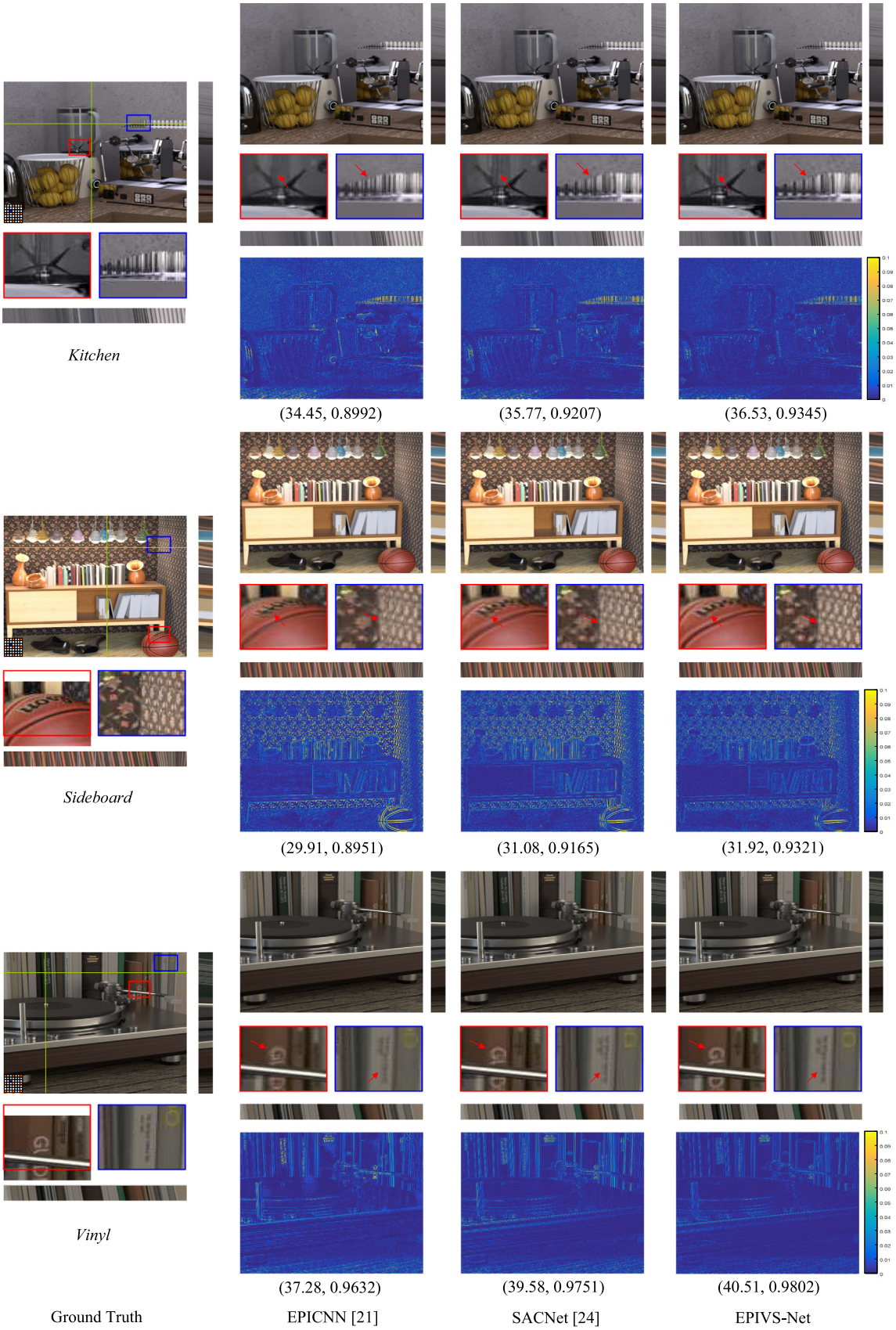

Fig. 9. Visual comparisons of reconstructed SAI for synthetic scenes using different methods under task $3 \times 3 \rightarrow 7 \times 7$. Comparison shows ground-truth SAIs, error maps of reconstructed SAIs in Y channel, close-up versions of SAI portions in blue and red boxes, and extracted EPIs located at green line.

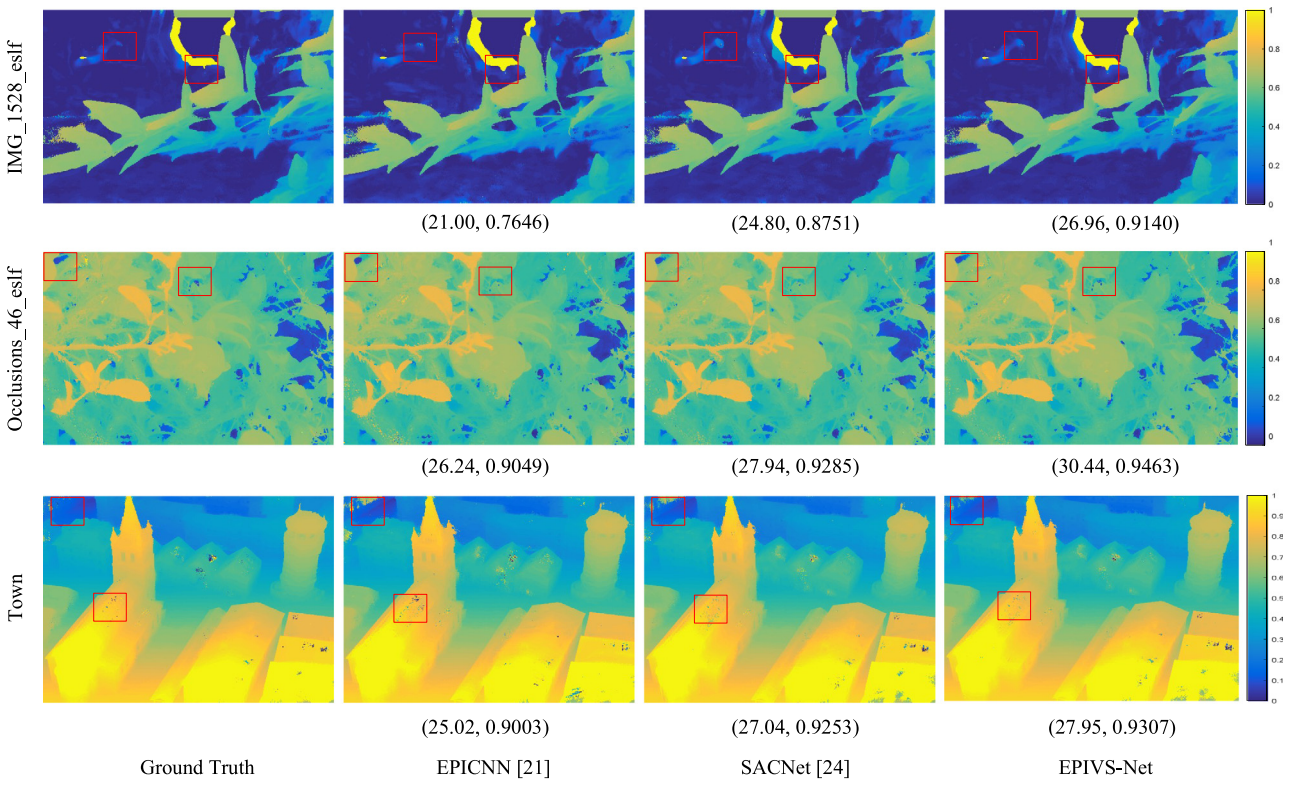

Fig. 10. Visual comparisons of depth maps estimated from the reconstructed 7 × 7 LF data by different methods. The depth estimation method in [46] is adopted to estimate the scene depth.

Table 6
Quantitative comparisons (PSNR/SSIM) of the three cases on real-world scenes under task 2 × 2 → 7 × 7.

	EPIVS-Net-H	EPIVS-Net-V	EPIVS-Net
30 Scenes	42.01/0.984	41.91/0.984	**42.14/0.985**
Reflective	40.39/0.969	40.26/0.968	**40.49/0.969**
Occlusions	36.80/0.941	36.70/0.940	**36.90/0.945**
Average	39.73/0.965	39.62/0.964	**39.84/0.966**

3.78 dB, 0.76 dB, 6.13 dB and 3.19 dB, respectively, in PSNR relative to LBVSNet [18], SACNet [22], EPICNN [11], and SEPINet [20]. The main reasons are as follows. Firstly, LBVSNet [18] synthesizes novel SAIs in a separate forward pass, where the LF structure is overlooked. Moreover, the disparity warping step influences the super-resolution quality, especially for cases of texture regions. Secondly, EPICNN [11] and SEPINet [20] all try to synthesize a high angular resolution LF by reconstructing each high-resolution EPI with only two rows or columns of pixels as inputs. It is difficult to retain the intrinsic structure of a reconstructed high angular resolution LF and restore the texture details with such a sparse input, especially for LF scenes with complex textures. Even though SEPINet [20] improves the network by fusing a set of sheared EPIs, the performance is still limited. Thirdly, SACNet [22] fails to fully explore the spatio-angular correlations in reconstruction process, which reduces the super-resolution quality. The proposed EPIVS-Net can perform better by making full use of spatial–angular clues, and make an interaction between the spatial and angular dimensions in the super-resolution process.

5.4. Ablation investigation

In order to further verify the influence of horizontal and vertical LF EPI-VSs to the super-resolution quality, we investigate the super-resolution performances with different EPI-VS inputs to our network under task 2 × 2 → 7 × 7. Three cases are under consideration: EPIVS-Net with only horizontal or vertical LF EPI-VS as input and EPIVS-Net with both horizontal and vertical LF EPI-VSs as inputs. For short, the method with only horizontal or vertical LF EPI-VS as input is referred to as EPIVS-Net-H or EPIVS-Net-V, respectively. And method with both horizontal and vertical LF EPI-VSs as inputs is also called EPIVS-Net.

Table 6 gives the quantitative comparisons of the three cases on real-world scenes under task 2 × 2 → 7 × 7. From Table 6, we find that the average performance is improved by using both horizontal and vertical LF EPI-VSs as inputs. The reason is that the spatio-angular correlations in both direction can be fully explored by using horizontal and vertical LF EPI-VS as inputs, which improves the reconstruction quality. The average performance of EPIVS-Net-H is a little higher than EPIVS-Net-V. But the gain is negligible. Note that, even though only using one directional LF EPI-VS as input, the super-resolution quality is still superior to other state-of-the-art methods, which can be seen from Tables 5 and 6. This further illustrates the superior of the proposed EPI-VS representation in LF angular super-resolution.

5.5. Computational time

Table 7 gives the execution time comparison of different methods in synthesizing one SAI. The compared methods are tested on a computer

Table 7
The execution time comparison of different methods in synthesizing one SAI. Running time in seconds are derived as the average execution time of all the test LF scenes on NVIDIA Quadro P5000 with GPU acceleration.

	LBVSNet [18]	EPICNN [11]	EP4DCNN [25]	P4DCNN [27]	EPIVS-Net
Execution time (s)	6.71	3.73	1.15	0.31	0.08

with Intel Xeon Gold 6150 @2.7 GHz, 180 GB RAM and NVIDIA Quadro P5000. From Table 7, we notice that LBVSNet [18] needs more execution time than the other methods. This is mainly because LBVSNet [18] can only reconstruct one SAI in one forward pass, and disparity and color estimation are also time-consuming. Compared to the EPICNN [11], EP4DCNN [25] and P4DCNN [27] need less execution time to synthesize one SAI. This is because EP4DCNN [25] and P4DCNN [27] can reconstruct more information in one forward pass based on EPI volume representation. Since the proposed EPIVS-Net can reconstruct all the SAIs in one forward pass, the proposed method need the least execution time than the other method. Moreover, in our method, the obtained horizontal and vertical feature maps after feature extraction sub-network are added together before being fed into the angular super-resolution network, which also decreases the computational burden.

5.6. Application for depth estimation

The LF data can reveal the geometry information of 3D scene, and can be used to estimate the scene depth. High quality LF data is advantageous in estimating scene depth map [46,47]. Therefore, in order to verify the proposed EPIVS-Net can restore more texture details during LF angular super-resolution, we compare the qualities of depth maps estimated from the reconstructed 7×7 LF data by different methods. A robust depth estimation method proposed in [46] is adopted to estimate the scene depth. Fig. 10 gives the visual comparisons of depth maps estimated from the reconstructed 7×7 LF data by different methods. From Fig. 10, we find that the proposed EPIVS-Net can achieve a better visual depth quality than EPICNN [11] and SACNet [22]. For example, for *IMG_1528_eslf* case, since the EPIVS-Net can preserve texture details in the reconstructed LF data, a high-quality depth map can be derived, shown in the red boxes in Fig. 10. The *Occlusions_46_eslf* contains complicated occlusions, which is challenging for depth estimation. The results by EPICNN [11] and SACNet [22] produce many errors in occlusion regions, which are alleviated by EPIVS-Net. This owes to that the EPIVS-Net can recover the occluded regions well and restore more texture details. The same results can also be obtained for *Town* case, which can validate the superiority of EPIVS-Net on preserving the texture information in reconstructed high angular resolution LF again.

5.7. Limitation

The main advantage of proposed method is that EPIVS-Net can implicitly learn LF SAI spatial correlation and angular correlation so as to recovery more texture details. However, for LF image with severe occlusion, the EPIVS-Net cannot produce high-quality results and achieve a high structural similarity of reconstructed SAIs, such as those shown in Fig. 11. This is because EPIVS-Net cannot precisely learn the spatial correlation and angular correlation in heavy occlusion areas and fail to keep LF similar structure.

6. Conclusion

In this paper, we propose a learning based dense LF angular super-resolution method based on LF EPI-VS representation. By learning a mapping from low angular resolution horizontal and vertical EPI-VS to high angular resolution LF EPI-VS, the proposed EPIVS-Net can reconstruct a high angular resolution LF effectively. In order to fully explore the LF spatio-angular correlations, we introduce a EPI-VS representation. And in the whole super-resolution procedure, 3D convolutions are adopted, which can accommodate the EPI-VS data well and allow information propagation between two spatial dimensions and one directional dimension of EPI-VS data. The overall EPIVS-Net contains four parts: feature extraction, angular super-resolution, texture refinement and integration, which allows the LF angular super-resolution in a coarse-to-fine manner.

Fig. 11. Failure cases on LF scene with severe occlusion.

The experimental results demonstrate that the proposed EPIVS-Net can synthesize high-quality high angular resolution LF data with low angular resolution SAI as inputs. It outperforms the state-of-the-art methods on synthetic and real-world LF scenes under two different tasks. Moreover, the application on depth estimation also verifies that the EPIVS-Net can recover more texture details, especially for some challenging LF scenes.

CRediT authorship contribution statement

Deyang Liu: Conceptualization, Methodology, Writing - original draft, Funding acquisition, Software. **Qiang Wu:** Supervision, Writing - review & editing. **Yan Huang:** Writing - review & editing. **Xinpeng Huang:** Supervision, Formal analysis, Resources, Funding acquisition. **Ping An:** Supervision, Formal analysis, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China, under Grants 61801006, 61828105, 62001279 and 62020106011, in part by China Postdoctoral Science Foundation under Grant 2021T140442.

References

[1] G. Wu, B. Masia, A. Jarabo, Y. Zhang, et al., Light field image processing: An overview, IEEE J. Sel. Top. Sign. Proces. 11 (7) (2017) 926–954.
[2] B. Wilburn, et al., High performance imaging using large camera arrays, ACM Trans. Graph. 24 (3) (2005) 765–776.
[3] J. Unger, A. Wenger, T. Hawkins, A. Gardner, P. Debevec, Capturing and rendering with incident light fields, in: Proc. Eurograph. Workshop Rendering Techn. 2003, pp. 141-149.
[4] Lytro, (2011). [Online]. Available: https://www.lytro.com/.
[5] RayTrix, (2010). [Online]. Available: http://www.raytrix.de/.
[6] C. Chen, Y. Lu, M. Su, Light field based digital refocusing using a DSLR camera with a pinhole array mask, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 754–757.
[7] J. Peng, Z. Xiong, Y. Wang, Y. Zhang, D. Liu, Zero-shot depth estimation from light field using a convolutional neural network, IEEE Trans. Comput. Imag. 6 (2020) 682–696.
[8] J. Chen, J. Hou, Y. Ni, L.P. Chau, Accurate light field depth estimation with superpixel regularization over partially occluded regions, IEEE Trans. Image Process. 27 (10) (2018) 4889–4900.

[9] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, M. Gross, Scene reconstruction from high spatio-angular resolution light fields, ACM Trans. Graph. 32 (4) (2013) 1–73.
[10] R.S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, P. Debevec, A system for acquiring, processing, and rendering panoramic light field stills for virtual reality, ACM Trans. Graph. 37 (6) (2018) 15, 197:1-197.
[11] G. Wu, Y. Liu, L. Fang, Q. Dai, T. Chai, Light field reconstruction using convolutional network on EPI and extended applications, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2019) 1681–1694.
[12] L. Shi, H. Hassanieh, A. Davis, D. Katabi, F. Durand, Light field reconstruction using sparsity in the continuous fourier domain, ACM Trans. Graph. 34 (1) (2014) 13, 12:1-12.
[13] K. Mitra, A. Veeraraghavan, Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 22–28.
[14] S. Vagharshakyan, R. Bregovic, A. Gotchev, Light field reconstruction using shearlet transform, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2018) 133–147.
[15] Z. Zhang, Y. Liu, Q. Dai, Light field from micro-baseline image pair, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3800–3809.
[16] F. Zhang, J. Wang, E. Shechtman, Z. Zhou, J. Shi, S. Hu, Plenopatch: Patch-based plenoptic image manipulation, IEEE Trans. Vis. Comput. Graphics 23 (5) (2017) 1561–1573.
[17] Y. Gao, R. Bregovic, A. Gotchev, R. Koch, MAST: Mask-accelerated shearlet transform for densely-sampled light field reconstruction, in: IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 187–192.
[18] N.K. Kalantari, T.C. Wang, R. Ramamoorthi, Learning based view synthesis for light field cameras, ACM Trans. Graph. 35 (6) (2016) 10, 193:1-193.
[19] A.K. Vadathya, S. Girish, K. Mitra, A unified learning-based framework for light field reconstruction from coded projections, IEEE Trans. Comput. Imag. 6 (2020) 304–316.
[20] G. Wu, Y. Liu, Q. Dai, T. Chai, Learning sheared EPI structure for light field reconstruction, IEEE Trans. Image Process. 28 (7) (2019) 3261–3273.
[21] N. Meng, H.K.H. So, X. Sun, et al., High-dimensional dense residual convolutional neural network for light field reconstruction, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1–14.
[22] W.F.H. Yeung, J. Hou, J. Chen, Y. Ying Chung, X. Chen, Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues, in: European Conference on Computer Vision (ECCV), 2018, pp. 137–152.
[23] J. Jin, J. Hou, H. Yuan, et al., Learning light field angular super-resolution via a geometry-aware network, in: 34th AAAI Conference on Artificial Intelligence, 2020, pp. 1–9.
[24] Z. Huang, J.A. Fessler, T.B. Norris, I.Y. Chun, Light-field reconstruction and depth estimation from focal stack images using convolutional neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 8648–8652.
[25] Y. Wang, F. Liu, K. Zhang, Z. Wang, Z. Sun, T. Tan, High-fidelity view synthesis for light field imaging with extended pseudo 4dcnn, in: IEEE Transactions on Computational Imaging, 2020, pp. 1–14.
[26] G. Wu, Y. Liu, L. Fang, T. Chai, Lapepi-net: A laplacian pyramid epi structure for learning-based dense light field reconstruction, 2019, [Online]. Available: https://arxiv.org/abs/1902.06221.
[27] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, T. Tan, End-to-end view synthesis for light field imaging with pseudo 4dcnn, in: European Conference on Computer Vision (ECCV), 2018, pp. 333–348.
[28] F. Battisti, M. Carli, P.L. Callet, A study on the impact of visualization techniques on light field perception, in: European Signal Processing Conference (EUSIPCO), 2018, pp. 2155–2159.
[29] R.A. Farrugia, C. Guillemot, A simple framework to leverage state-of-the-art single-image super-resolution methods to restore light fields, Signal Process., Image Commun. 80 (2020) 115638.
[30] X. Li, G. Cao, Y. Zhang, A. Shafique, P. Fu, Combining synthesis sparse with analysis sparse for single image super-resolution, Signal Process., Image Commun. 83 (2020) 115805.
[31] Z. Cheng, Z. Xiong, D. Liu, Light field super-resolution by jointly exploiting internal and external similarities, IEEE Trans. Circuits Syst. Video Technol. (2019) 1–13.
[32] S. Zhang, Y. Lin, H. Sheng, Residual networks for light field image super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11046–11055.
[33] J. Zhang, M. Shao, L. Yu, Y. Li, Image super-resolution reconstruction based on sparse representation and deep learning, Signal Process., Image Commun. 87 (2020) 115925.
[34] C. Michael, G.J. Steven, S. Richard, G. Radek, S. Rick, The lumigraph, in: SIGGRAPH, 1996.
[35] C. Zou, X. Huang, Hyperspectral image super-resolution combining with deep learning and spectral unmixing, Signal Process., Image Commun. 84 (2020) 115833.
[36] W. Lai, J. Huang, N. Ahuja, M. Yang, Fast and accurate image super-resolution with deep Laplacian pyramid networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2019) 2599–2613.
[37] D. Xiong, Q. Gui, W. Hou, M. Ding, Gradient boosting for single image super-resolution, Inform. Sci. 454–455 (2018) 328–343.
[38] J. Jin, J. Hou, J. Chen, et al., Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion, 2020, [Online]. Available: https://arxiv.org/abs/1909.01341.
[39] A. Vedaldi, K. Lenc, Matconvnet C convolutional neural networks for matlab, in: ACM International Conference on Multimedia, 2015, pp. 689–692.
[40] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.
[41] P. Benecki, M. Kawulok, D. Kostrzewa, L. Skonieczny, Evaluating super-resolution reconstruction of satellite images, Acta Astronaut. 153 (2018) 15–25.
[42] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3147–3155.
[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[44] K. Honauer, O. Johannsen, D. Kondermann, B. Goldluecke, A dataset and evaluation methodology for depth estimation on 4D light fields, in: Asian Conference on Computer Vision, 2016, pp. 19–34.
[45] A.S. Raj, M. Lowney, R. Shah, G. Wetzstein, Stanford lytro light field archive, 2016, http://lightfields.stanford.edu/LF2016.html, [Online].
[46] S. Zhang, H. Sheng, C. Li, J. Zhang, X. Zhang, Robust depth estimation for light field via spinning parallelogram operator, Comput. Vis. Image Understand. 145 (2016) 148–159.
[47] H. Sheng, P. Zhao, S. Zhang, J. Zhang, D. Yang, Occlusion-aware depth estimation for light field using multi-orientation EPIs, Pattern Recognit. 74 (2018) 587–599.