Pattern Recognition 140 (2023) 109551

Contents lists available at ScienceDirect

Pattern Recognition

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

Dense light field reconstruction based on epipolar focus spectrum

Yaning Li[a], Xue Wang[a], Hao Zhu[b], Guoqing Zhou[a], Qing Wang[a,*]

[a] School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China
[b] School of Electronic Science and Engineering, Nanjing University, Nanjing, 210023, China

ARTICLE INFO

Article history:
Received 15 May 2022
Revised 3 January 2023
Accepted 21 March 2023
Available online 23 March 2023

Keywords:
Light field representation
Epipolar focus spectrum (EFS)
Dense light field reconstruction
Depth independent
Frequency domain

ABSTRACT

Existing light field (LF) representations, such as epipolar plane image (EPI) and sub-aperture images, do not consider the structural characteristics across the views, so they usually require additional disparity and spatial structure cues for follow-up tasks. Besides, they have difficulties dealing with occlusions or large disparity scenes. To this end, this paper proposes a novel Epipolar Focus Spectrum (EFS) representation by rearranging the EPI spectrum. Different from the classical EPI representation where an EPI line corresponds to a specific depth, there is a one-to-one mapping from the EFS line to the view. By exploring the EFS sampling task, the analytical function is derived for constructing a non-aliasing EFS. To demonstrate its effectiveness, we develop a trainable EFS-based pipeline for light field reconstruction, where a dense light field can be reconstructed by compensating the missing EFS lines given a sparse light field, yielding promising results with cross-view consistency, especially in the presence of severe occlusion and large disparity. Experimental results on both synthetic and real-world datasets demonstrate the validity and superiority of the proposed method over SOTA methods.

1. Introduction

Light field [1] imaging system records the 3D scene in both spatial and angular domains [2,3], and has becoming one of the most potential techniques for immersive virtual reality [4,5]. However, due to the spatio-angular trade-off [6] in the sampling process, it is expensive to acquire high angular resolution light fields (LFs), which limits the application and development of light field technologies. Light field reconstruction aims at synthesizing LFs from sparse input views and serves as an essential tool for generating dense LFs.

In decades, dense light field reconstruction has drawn a lot of attention and gained great progress, however it still faces many challenging issues. For depth-based and optical flow-based methods [7–10], the reconstruction results are prone to depth estimation and the view consistency could not be preserved well. For implicit depth-based methods, *i.e.*, the multiplane image (MPI) representation [11], the additionally introduced transparency term could not describe intricately occluded areas well (see Figs. 10, and 11).

Since the essence of dense light field reconstruction is to eliminate the aliasing contents in the Fourier spectrum of the angularly undersampled light field [12], several methods have been recently proposed to focus on recovering the high-frequency spectrum either by modeling the texture consistency in the spatial domain [13] or inpainting in the transformed domain [14,15]. However, due to the information asymmetry [16] between the spatial and angular dimensions, the high-frequency spectrum learned or modeled from the LFs with small disparities is inapplicable to reconstruct the LFs with large disparities, causing artifacts near the occlusion boundaries.

More recently, Li et al. [17] propose to eliminate the aliasing contents in the refocused images using the 2D focal stack spectrum. The depth-independent property of line distribution in the 2D focal stack spectrum provides the theory basis for performing a unified anti-aliasing rendering for all depth layers. However, due to the irreversibility of the 'integral' operator, this learning framework could only provide anti-aliasing results for the focal stack and it still fails to recover high angular resolution LFs.

In this paper, we extend the theory of the 2D focal stack spectrum. We observed that a 2D focal stack spectrum could be directly obtained by applying the 1D Fourier transform to a re-arranged EPI spectrum. Since the 'Fourier transform' and 're-arrange' operations are both reversible, a complete loop could be established between an EPI and a 2D focal stack spectrum, hence the task of reconstructing high angular resolution LFs could be tackled as the spectrum completion problem. To better represent the connection between the 2D focal stack spectrum and the EPI, the term '2D focal stack spectrum' is renamed as 'Epipolar Focus Spectrum' (EFS, in Section 3) in the paper. In addition, to eliminate the aliasing

* Corresponding author.
E-mail address: qwang@nwpu.edu.cn (Q. Wang).

https://doi.org/10.1016/j.patcog.2023.109551
0031-3203/

caused by insufficient focal stack layers, we analyze the problem of EFS sampling and derive the analytical function involving the minimal focal stack layers, the scene distribution function, camera parameters, the number of views, and scene depth. Based on the depth-independent property of the EFS, it is possible to pursue a unified light field reconstruction to process full disparity contents simultaneously. We first present an end-to-end convolutional neural network (CNN) to eliminate the aliasing contents in the EFS formed from an undersampled light field (in Section 4). Then the generated non-aliasing EFS is projected to construct the EPI spectrum. After applying the inverse Fourier transform (IFT) to obtain the dense light field, a U-Net with a perceptual loss is finally utilized to optimize the reconstructed result and eliminate the 'trailing image' [18] caused by the integral operation, especially in the marginal views. Experimental results (in Section 5) verify the effectiveness of the proposed EFS-based dense light field reconstruction method.

The main contributions of the work include,

1) The theory of EFS is improved in two aspects. a) The complete reversible loop between the EPI and the EFS is established. b) The EFS sampling is modeled using an analytical function. As a result, the cross-view consistency is guaranteed for full depth/disparity light field reconstruction.

2) An EFS-based learning framework for dense light field reconstruction is proposed. Extensive experiments on both synthetic and real light field datasets verify the superiority of the proposed method.

2. Related work

2.1. Light field representation

Let $L(u, v, x, y)$ represent the distribution of rays in 3D space, where (u, v) and (x, y) denote the intersections between the ray with angular/camera and spatial/image planes, respectively [2,3]. To better model the contents, several representations have been proposed in the literature. In the spatial domain, the sub-aperture image and EPI are the two most commonly used representations. The former emphasizes the spatial information per view, while the latter focuses on the disparity among views, where the slope of the EPI line is associated with the disparity/depth. In the Fourier domain, by exploring the equivalence between the 3D focal stack and a 4D light field, Ng [19] claim the 2D spectrum of a refocused image could be obtained by slicing the corresponding 4D spectrum. Dansereau et al. [20] propose the hyper-cone and hyper-fan representations to extend the focal range of each focal slice. Le Pendu et al. [21] analyse the sparsity of light field spectrum and propose the Fourier disparity layer (FDL) representation by assuming the spectrum energy concentrates on several slices.

Since these representations are highly correlated to the scene depth, directly applying the features extracted from the LFs with small disparity range to the LFs with large disparity range might lead to wrong inference. In contrast, our proposed depth-invariant EFS representation can enable an operation or processing covering the whole depth range. It is necessary to apply a depth-invariant representation for light field processing within the whole depth range.

2.2. Anti-aliasing of refocusing

Insufficient angular sampling results in aliasing in the refocused images. Researchers have proposed many anti-aliasing solutions in both spatial and frequency domains. In the spatial domain, Levoy and Hanrahan [2] prefilter the light field to reduce the aliasing. Chang et al. [22] compensate the effect of undersampling by utilizing depth information. Xiao et al. [23] first detect the aliasing contents by analyzing the angular aliasing model in the spatial domain. According to this model, the aliasing could be removed as the lower-frequency terms of the decomposition at the refocusing stage. In the frequency domain, Isaksen et al. [24] dynamically reparameterize the light field, allowing exact spectrum recovery of a single point without post-aliasing. Chai et al. [12] analyse the trade-off between sampling density and depth resolution. Based on focal stacks and sparse collections of viewpoints, Levin and Durand [25] employ the focal manifold in derivations of 2D deconvolution kernels [21]. After that, Lumsdaine et al. [26] discuss the aliasing in terms of the focal manifold and conclude by rendering wide depth-of-field images. By deriving the frequency domain of support of the light field, Dansereau et al. [20] present a simple, linear single-step filter to achieve volumetric focus effects.

All these methods consider the 3D focal stack as multiple individual slices and remove slice-wise aliasing contents, thus the consistency between neighboring slices is not preserved. Li et al. [17] propose an anti-aliasing method by completing the focal stack spectrum, where the aliasing contents of all the slices are handled at the same time by treating the 3D focal stack as a whole and the cross-view consistency could be well maintained.

2.3. Dense reconstruction

As mentioned above, the anti-aliasing operation essentially corresponds to the super-resolution operation in the angular domain [8]. Existing angular super-resolution methods could be mainly divided into two categories.

The first category is based on depth estimation [7–10,27,28]. Wanner and Goldluecke [7] reconstruct novel views by combining input views and estimated depth information. Kalantari et al. [8] propose two convolutional neural networks to estimate depth and color of each viewpoint sequentially. Srinivasan et al. [9] take one 2D RGB image as input and synthesize a 4D RGBD light field. Specifically, their pipeline consists of one CNN that estimates scene geometry, and another CNN that predicts occluded rays and non-Lambertian effects. Subsequently, Srinivasan et al. [27] propose to utilize the MPI representation to synthesize the viewpoint from a narrow baseline stereo pair. Liu et al. [10] extend the traditional 2D optical flow model to 4D and realize dense light field reconstruction by calculating 4D light field flow. DILF [28] takes an optical flow as input and proposes a learnable model, namely dynamic interpolation, to replace the commonly-used geometry warping operation for novel view generation.

The second category focuses on modeling the consistency of EPI texture [13,16,29,30] or the sparsity of Fourier spectrum [14,15,25]. Wu et al. [13] convert the light field reconstruction task to a one-dimensional super-resolution of the 2D EPI. Zhu et al. [30] improve the super-resolution performance on EPI in large disparity areas by introducing a long-short term memory module. Considering the special 2D mesh sampling structure of the 4D light field, Levin and Durand [25] utilize the 3D focus stack to complement the spectrum of the 4D light field. Shi et al. [14] exploit the sparsity of the 4D light field in the continuous Fourier domain and perform dense reconstruction by adopting the sparse Fourier transform. Vagharshakyan et al. [15] utilize a sparse representation of underlying EPIs in the shearlet domain and employ an iterative regularized reconstruction.

Recently, neural rendering [31] has achieved great success in 3D vision. Mildenhall et al. [32] model the rays as an implicit neural radiance field (NeRF) with a 5D input (3D for position and 2D for view angles) and a 4D output. Attal et al. [33] improve NeRF by replacing the 5D input with the 4D coordinates of the ray and achieve state-of-the-art results in forward-facing datasets. However, these methods require a large number of input views for training and most of them are not generalizable models, which

Table 1
Related notations of the EFS representation.

Term	Definition
$L(u, v, x, y)$	A 4D light field
u, v	Angular coordinates
x, y	Spatial coordinates
$E(u, x)$	2D EPI
$E_d(u, x)$	Sheared EPI at the specific disparity d
d_{range}	Disparity range for the shearing process
N_f	The number of refocus layers
$FT_*(\cdot)$	1D Fourier transform on the variable $*$
$FT_{2D}(\cdot)$	2D Fourier transform
$F(f, x)$	Focal stack integrated by $E_d(u, x)\|_{d=f}$
$\mathcal{E}(\omega_u, \omega_x)$	Fourier spectrum of $E(u, x)$
$\mathcal{F}(f, \omega_x)$	Slicing and rearranging of $\mathcal{E}(\omega_u, \omega_x)$
$EFS(\omega_f, \omega_x)$	EFS or 2D Fourier spectrum of $F(f, x)$
u_{ref}	Reference view (the center view in this work)

means it is essential to retrain the network (often taking several hours) for different scenes.

Nevertheless, these methods either rely on accurate depth calculations or are inappropriate for large disparity scenes since the required texture lines or sparse spectrum features are not available. Differently, based on the depth-independent EFS representation, our proposed method skips the challenging depth estimation and optimizes the contents at various depths with the same strategy, especially in the presence of severe occlusion and large disparity (please refer to the supplementary material for theoretical discussion and analysis on occlusion and large disparity).

3. The EFS representation

In this section, we first define the EFS representation for a light field based on [17] and then introduce its depth-independent characteristics.

3.1. Notations

For better understanding the definition of EFS, we first list the notations used in this work in Table 1.

Given a 4D light field $L(u, v, x, y)$, where (u, v) and (x, y) refer to the angular and spatial dimensions respectively. $E(u, x)$ denotes the particular EPI where $v = v^*$ and $y = y^*$. $E_d(u, x)$ is the sheared EPI using the shearing operation $E_d(u, x) = E(u, x + d \cdot (u - u_{ref}))$ where u_{ref} represents the reference view. $F(f, x)$ denotes the 2D focal stack integrated by $E_d(u, x)$ when the sheared value $d = f$. $FT_*(\cdot)$ and $FT_{2D}(\cdot)$ denote the 1D and 2D Fourier transform respectively ($*$ means a specific variable). $\mathcal{E}(\omega_u, \omega_x)$ is the Fourier spectrum of $E(u, x)$. $EFS(\omega_f, \omega_x)$ is the 2D focal stack spectrum or the Epipolar Focus Spectrum (EFS) of $E(u, x)$.

3.2. Background

Given a 2D EPI $E(u, x)$, its EFS representation could be constructed by applying the shearing, integral and 2D Fourier transform operators to $E(u, x)$ successively (see the top row of Fig. 1),

$$E_f(u, x) = E(u, x + f \cdot (u - u_{ref})), \tag{1a}$$

$$F(f, x) = \int E_f(u, x) dx, \tag{1b}$$

$$EFS(\omega_f, \omega_x) = FT_{2D}(F(f, x)). \tag{1c}$$

According to Li et al. [17], the EFS is composed of multiple lines passing through the origin. Each line corresponds to a specific view in $E(u, x)$. The slope of each line is determined by the view index, the reference view index and the shearing step $\Delta\alpha$. It is worth noting that the slope of each view is independent from the scene depth. Additionally, the EFS is conjugate symmetric according to the property of the Fourier transform [34]. Please refer to [17] and Fig. 2 for more details.

3.3. EFS in the Fourier domain

The EFS could also be constructed directly in the Fourier domain (see the bottom row of Fig. 1),

$$\mathcal{E}(\omega_u, \omega_x) = FT_{2D}(E(u, x)) \tag{2a}$$

$$\mathcal{F}(f, \omega_x) = \mathcal{E}(-f\omega_x, \omega_x) \tag{2b}$$

$$EFS(\omega_f, \omega_x) = FT_f(\mathcal{F}(f, \omega_x)). \tag{2c}$$

It is straightforward to prove the equivalence between the constructions in the Fourier domain (Eq. (2)) and in the spatial domain (Eq. (1)). According to the Fourier slice photography theory [19], the Fourier spectrum of the line f in the focal stack $F(f, x)$ is equal to slicing the spectrum of EPI $\mathcal{E}(u, x)$ with slope $-f$, *i.e.*, Eq. (2b). Therefore, Eq. (2b) is equal to applying a 1D Fourier transform to the focal stack along the x-dimension. Because both x and f dimensions of a focal stack are Fourier transformed in the EFS construction (Eq. (1)), it is necessary to apply a 1D Fourier transform to $\mathcal{F}(f, \omega_x)$ via Eq. (2c).

Noting that, the lossless EFS could only be constructed in two cases, *i.e.*, the range of the slicing operation in Eq. (2c) meets $\{f_{\min}, f_{\max}\} \to \{-\infty, +\infty\}$ or with an infinite aperture size [25]. However, since these conditions are practically impossible to achieve, the EFS is actually a lossy representation in practice. To minimize the effects of the missing spectrum in $\mathcal{E}(-f\omega_u, \omega_x)$, it is suggested to set $d_{\min}$ and $d_{\max}$ as the minimum and maximum disparities of the scene respectively. A detailed analysis on this issue is provided in Section 5.3.

3.4. Sampling analysis of EFS

As analysed above, the aliasing occurs on the focal stack when the view is undersampled in the spatial domain (Fig. 2(c)). The spectrum lines of EFS will also become discrete (Fig. 2(d)) in the frequency domain. We complete the EFS by the method proposed in Section 4.1 to remove this aliasing. In addition, the insufficient refocus layers also cause aliasing in EFS (Fig. 3). Here we focus on the second case caused by insufficient refocus layers and analyse the lower bound of focal stack layers required for non-aliased EFS recovery.

As shown in Fig. 3, when the refocus range of the focal stack is equal to the depth range for all the objects in the scene, aliasing appears when the disparity gap $\Delta\alpha$ between neighboring focal layers increases. To eliminate aliasing, all lines formed from different views in the focal stack ought to be continuous instead of being discrete, *i.e.*, $\Delta\alpha(u_i - u_{ref}) \le 1,\ \forall i \in [1, N_u]$. All the inequalities for the middle views hold if the inequalities for two marginal views hold. In other words, the disparity gap $\Delta\alpha$ should meet the following inequality (please refer to the supplementary material for more theoretical explanation of Eq. (3)),

$$\Delta\alpha \frac{N_u - 1}{2} \le 1. \tag{3}$$

The refocus range $[d_{\min}, d_{\max}]$ is set according to the depth range $[Z_{\min}, Z_{\max}]$,

$$d_{\max} = \frac{kB}{Z_{\min}},\ d_{\min} = \frac{kB}{Z_{\max}}, \tag{4}$$

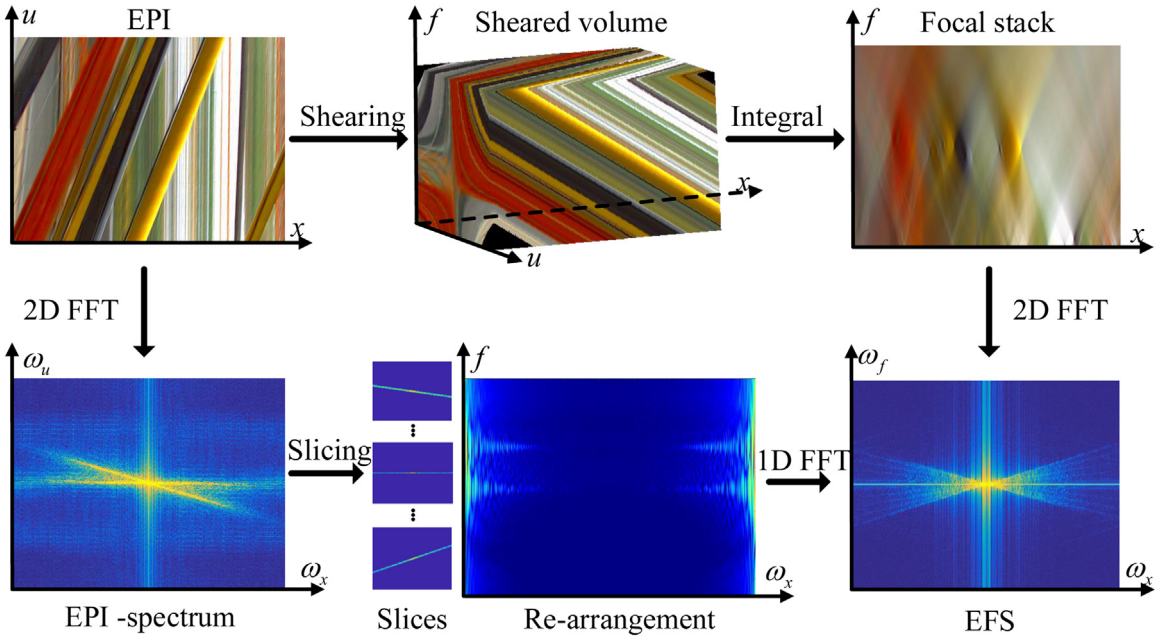


Fig. 1. Two different ways to obtain the EFS, either via the focal stack (top flow) or via the EPI spectrum (bottom flow).

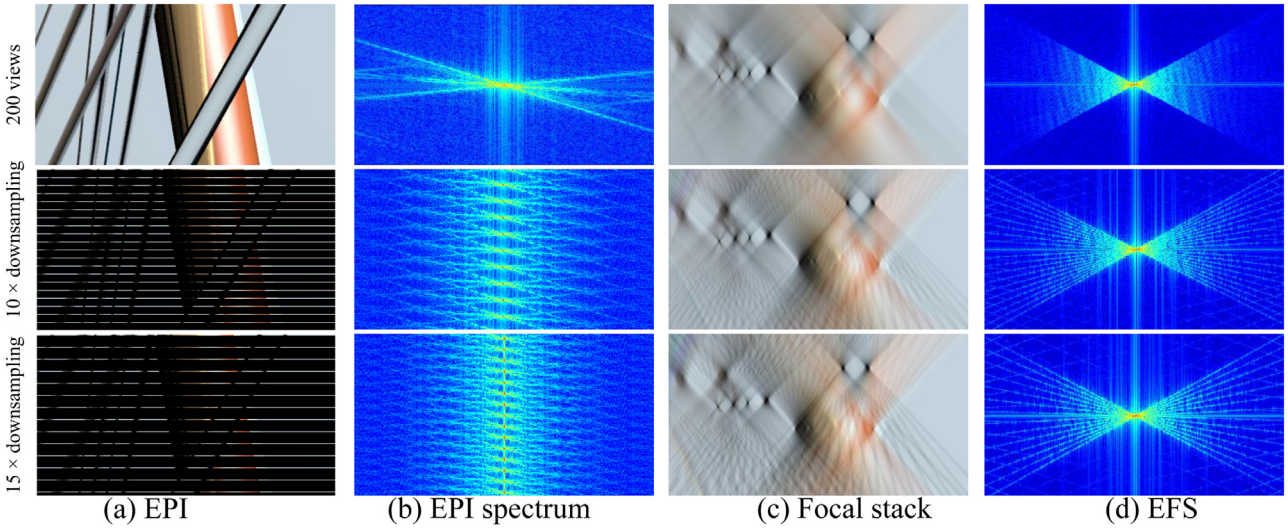


Fig. 2. Illustrations of EPI and EFS under different angular sampling rates. From TOP to BOTTOM: 200 views (original), 10× downsampling and 15× downsampling respectively.

where k is the focal length and B is the baseline. Thus, the number of refocus layers N_f in the focal stack is calculated by

$$N_f = \frac{d_{\max} - d_{\min}}{\Delta\alpha} = \frac{kB(Z_{\max} - Z_{\min})}{\Delta\alpha(Z_{\min}Z_{\max})}. \tag{5}$$

Combining Eqs. (3) and (5), we have

$$N_f \geq \frac{kB(Z_{\max} - Z_{\min})(N_u - 1)}{2Z_{\max}Z_{\min}}. \tag{6}$$

When the depth is discontinuous, it is essential to take the scene distribution into consideration. The minimum number of focal layers is estimated as

$$N_{fmin} = S(Z, O, T)\frac{kB(N_u - 1)\Delta Z}{2(Z_{\min} + \Delta Z)Z_{\min}}, \tag{7}$$

where $\Delta Z = Z_{\max} - Z_{\min}$, $S(Z, O, T)$ denotes the scene distribution function, determined by the depth Z, the occlusion O and the texture T. (Please refer to the supplementary material for more detailed discussion of the distribution function $S(Z, O, T)$).

In summary, the lower bound of focal stack layers is determined by the relative depth variation $(Z_{\min}, Z_{\max})$, the scene distribution $S(Z, O, T)$ and the light field camera parameters kB. Fig. 4 illustrates an example of choosing N_{fmin}, where $N_u = 200$, $S(Z, O, T) = 1$, $kB = 9$, $Z_{\min} \in [2, 10]$ (varying the minimum depth to model the scene at different distances), $\Delta Z \in [1, 100]$. Particularly, the black curve shows the varying trend of N_{fmin} with $Z_{\min} = 4$. Please refer to Section 5.3 for more experimental analysis.

4. EFS-based dense reconstruction

Insufficient angular sampling causes aliasing in the focal stack, thus reconstructing a dense light field is equivalent to restoring a complete EFS corresponding to the non-aliasing light field/focal stack. As analysed in Section 3.2 and [17], we can complement the non-aliased EFS by learning to recover the spectrum lines corresponding to missing viewpoints. Therefore, the light field reconstruction task could be decomposed into three sub-tasks, *i.e.*, EFS

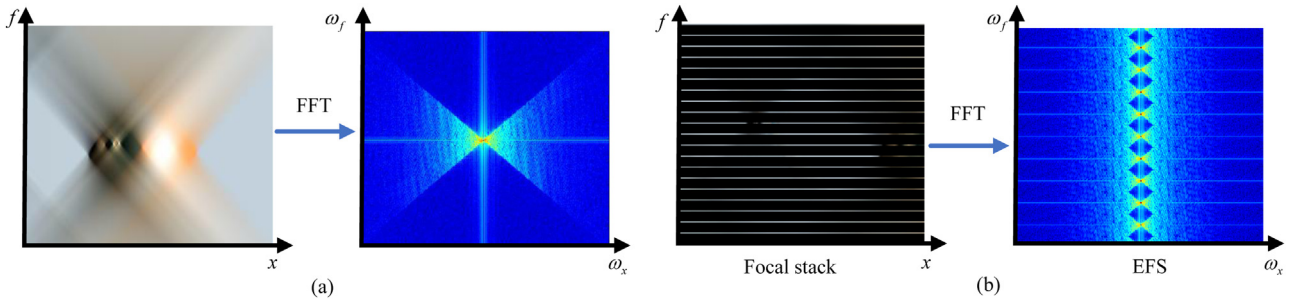


Fig. 3. Illustration of EFSs under different focal layer counts. (a) Sufficient focal layers result in a non-aliased EFS. (b) Insufficient focal layers result in an aliased EFS.

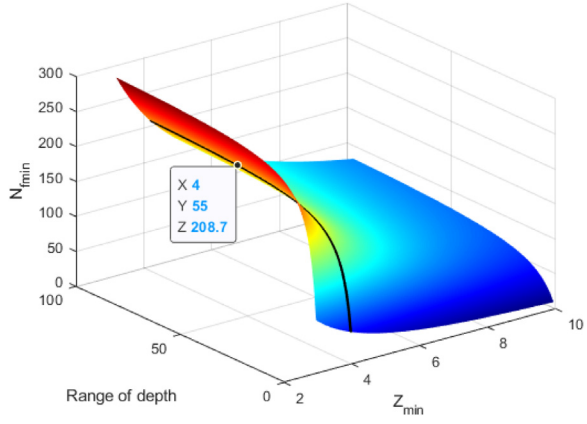


Fig. 4. The varying trend of N_{fmin} regarding $Z_{\min}$ and depth range.

reconstruction, EPI spectrum reconstruction and EPI optimization in the spatial domain. For the EFS reconstruction task, the method presented in [17] is adopted. The pipeline of the proposed dense light field reconstruction is shown in Fig. 5.

4.1. EFS reconstruction

Specifically, similar to Li et al. [17], we first perform shearing on an undersampled EPI to get the focal stack via Eq. (1a), then apply the Fourier transform on the aliased focal stack to get the aliased EFS via Eq. (1b). Subsequently, the loss function *loss* for optimizing the aliased EFS is

$$loss = \|\phi_\sigma(EFS_{ali}) - EFS_{gt}\|_2 + \lambda loss_s, \tag{8}$$

where ϕ represents the CNN parameterized by σ, the scalar λ is set to 1.5 for balancing the contributions of the two loss terms. The second term $loss_s$ constrains the conjugate symmetry of the reconstructed EFS,

$$loss_s = \frac{1}{N_f W}\sum_{i=0}^{N_f-1}\sum_{j=0}^{W-1}\left|EFS(\omega_i,\omega_j) - EFS^*(-\omega_i,-\omega_j)\right|, \tag{9}$$

where $|\cdot|$ refers to the norm of a complex number and $*$ indicates the standard conjugate operation. N_f is the number of refocus layers and W is the width of the sub-aperture image.

Two neural networks are used to extract features from the power spectrum and the phase angle, respectively. Then these features are combined using the Euler's formula to obtain the real and imaginary parts, which are concatenated and passed into the CNN layers for optimization. Please refer to [17] for more details.

4.2. EPI spectrum reconstruction

The Fourier Slice Imaging Theorem [19] tells a 2D slice through the origin of a 4D light field spectrum corresponds to a refocused image at a certain depth in the frequency domain. Based on this, we first apply the 1D inverse Fourier transform (IFT) of EFS along the f-axis,

$$\mathcal{F}(f,\omega_x) = \frac{1}{N_f}\sum_{\omega_f=0}^{N_f-1} EFS(\omega_f,\omega_x)e^{j2\pi\frac{f\omega_f}{N_f}}. \tag{10}$$

The projection relationship between $\mathcal{F}(f,\omega_x)$ and $\mathcal{E}(\omega_u,\omega_x)$ can be obtained via Eq. (2). Thus the Fourier spectrum $\mathcal{E}_{efs}(\omega_u,\omega_x)$ of the reconstructed EPI could be calculated by performing a reverse projection,

$$\mathcal{E}_{efs}(\omega_u,\omega_x) = \mathcal{F}\left(-\frac{\omega_u}{\omega_x},\omega_x\right). \tag{11}$$

Fig. 6 shows the diagram of this reverse projection. Due to the limitation of the disparity range, only the spectrum labeled as purple of Fig. 6(b) can be reconstructed using this operation.

4.3. EPI optimization

The next step is to apply the 2D IFT to get $E_{efs}(u,x)$ from $\mathcal{E}_{efs}(\omega_u,\omega_x)$. Since the interpolation operation is used during the focal stack construction, the 'tailing' effect appears after the inverse Fourier slice operation, especially for the marginal views which are far away from the reference view. Hence we use the U-Net Ψ_μ with a perceptual loss to optimize the reconstructed EPIs E_{efs},

$$\arg\min_{\mu}\left\{\left\|E_{gt},\Psi_\mu(E_{efs})\right\|\right\}, \tag{12}$$

where E_{gt} represents the ground truth for EPI.

The loss function for optimization is defined as follow,

$$loss_{EPI} = loss_{MAE} + \gamma_1 loss_{SSIM} + \gamma_2 loss_{VGG}, \tag{13}$$

where $loss_{MAE}$ is the Mean Absolute Error loss, $loss_{SSIM}$ is the Structural Similarity loss [35], and $loss_{VGG}$ is the perceptual loss [36] which is based on the VGG19 network trained on ImageNet. The scalars γ_i $(i=1,2)$ are set to 3 and 5 for balancing the effects of different loss terms. The detailed structure of this network is illustrated in the supplementary material.

The complete dense light field reconstruction algorithm is given in Algorithm 1. H represents the height of the sub-aperture image.

5. Evaluations

To evaluate the proposed method, we conduct experiments on both synthetic and real light field datasets. The real light field datasets are captured by both the camera array and the Lytro Illum camera. Four SOTA methods are compared, including Wu et al. [13] (without depth), LLFF [37] (MPI-based), DILF [28] (with depth) and NeLFRSE [33] (NeRF-based). LLFF is retrained on our dataset using the authors' released code for a fair comparison. Since Wu et al. [13] do not provide the training code, we use the pretrained model provided by the authors.

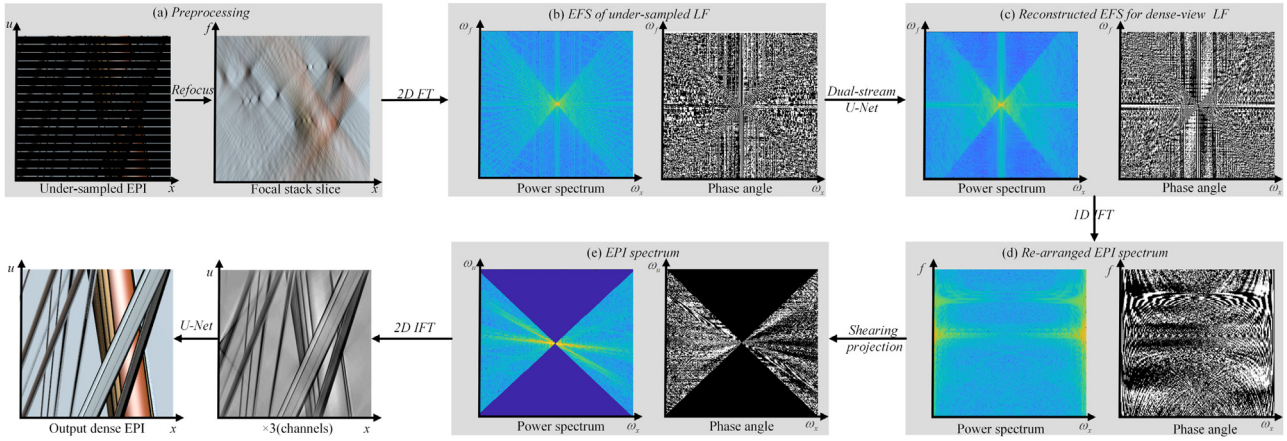


Fig. 5. The pipeline of the proposed EFS-based dense light field reconstruction, including preprocessing, EFS reconstruction, shearing projection and final optimization of the reconstructed EPI.

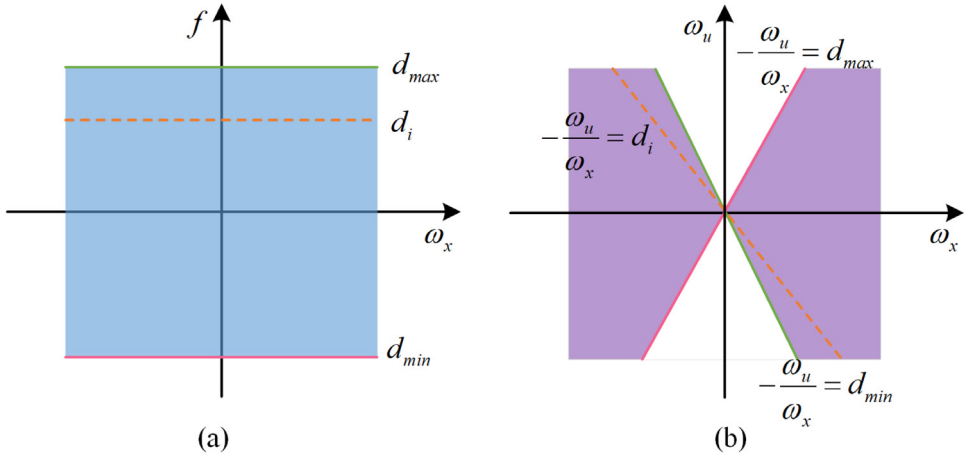


Fig. 6. The diagram of the projection from Fig. 5(d) to (e). (a) The 1D IFT of EFS along the f-axis $\mathcal{F}(f, \omega_x)$ (Fig. 5(d)). (b) The reconstructed EPI spectrum $\mathcal{E}_{efs}(\omega_u, \omega_x)$ (Fig. 5(e)). $[d_{\min}\ d_{\max}]$ is the disparity (refocus) range of the scene.

Algorithm 1

Input:

An undersampled light field with the disparity range d_{range}. The number of EFS layers N_f.

Output:

The reconstructed dense light field.

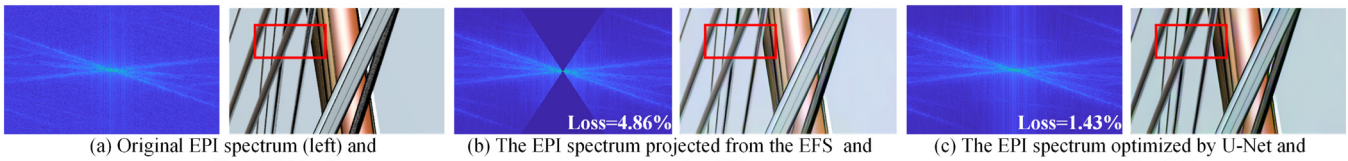
for $i = 1$ to H **do**
  Get the EPI $E(u, x)$.
  Perform the shearing operation on $E(u, x)$ within d_{range} via Eqn.1a.
  Get the aliased EFS_{ali} via Eqn.1b.
  Reconstruct the non-aliased $EFS(\omega_f, \omega_x)$ using the dual-stream U-Net ϕ (as shown in[17]).
  Perform 1D IFT on $EFS_{inv}(\omega_f, \omega_x)$ via Eqn.10.
  Reconstruct $\mathcal{E}_{efs}(\omega_u, \omega_x)$ via Eqn.11.
  Perform 2D IFT on $\mathcal{E}_{efs}(\omega_u, \omega_x)$ to get E_{efs}.
  Optimize the reconstructed EPI with the U-Net Ψ.
end for
Output the reconstructed dense light field.

To empirically validate the robustness of the proposed method, we perform evaluations under different downsampling patterns. The quantitative evaluations are performed by measuring the average PSNR and SSIM metrics over the synthetic views of the luminance channel. We also analyse the spectrum energy losses for EFS reconstruction and EPI reconstruction respectively.

5.1. Datasets and implementation details

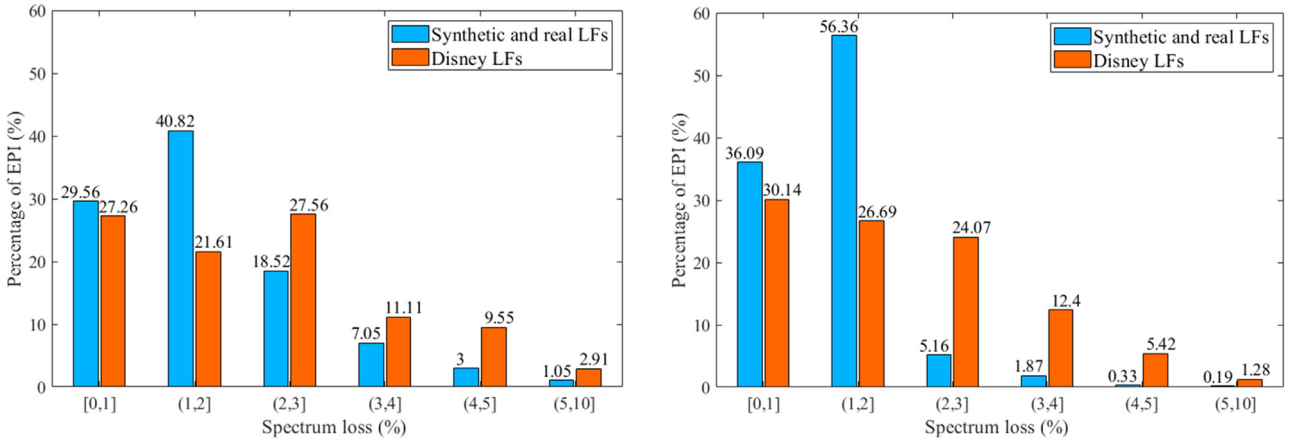
In the training process, both the synthetic and real LFs are used. For the synthetic data [38], 12 LFs containing complex textured structures are rendered using the automatic light field generator [30], of which 7 are for training and 5 for testing. Real LFs [38] are taken from the high-resolution Lytro Illum dataset [29], of which 20 are for training and 6 for testing. In order to show the relationship between viewpoints and EFS lines, we utilize the first 200 viewpoints for experiments. Additionally, the LFs from the Disney [39] dataset are used to verify the performance of the proposed method on unseen scenes captured by a camera array. For the dense LFs, the disparity between two adjacent views is less than one pixel for most scenes, and reaches two pixels for few scenarios.

At present, only the disparity along one single direction is concerned so that the 2D EPIs can represent the input light field. For each light field in the experiment, by considering the disparity and the scene distribution, the EFS is constructed by performing the refocus operation for 200 times ($N_f = 200$) where $\Delta\alpha = 0.01$. The scene disparity range determines the refocusing operation range f. For the details of the parameters of all the datasets, please refer to the supplementary material.



(a) Original EPI spectrum (left) and corresponding EPI (right)
(b) The EPI spectrum projected from the EFS and corresponding EPI
(c) The EPI spectrum optimized by U-Net and corresponding EPI

Fig. 7. Comparison of the EPI and its spectrum at different stages.



(a) The EPI spectrum loss without U-Net optimization
(b) The EPI spectrum loss with U-Net optimization

Fig. 8. Distributions of EPI spectrum loss on several light field datasets.

5.2. Spectrum domain

EPI spectrum loss analysis. In our experiments, the disparity range $d_{range} = [d_{\min}, d_{\max}]$ is also the refocus range of the focal stack. With the infinite aperture assumption, the spectrum energy could be regarded as zero for other focus layers beyond the constructed focal stack [25]. However, since the infinite aperture camera is currently not available, the EFS representation for a light field has a certain loss. Fig. 7 compares the spectra and EPIs recovered from the EFS and after the final optimization with the ground truth. Although the structural information in the EPI can be reconstructed (the disparity range is $[-1, 1]$ in the focal stack), the missing spectrum results in an uneven color distortion (Fig. 7(b)). The proposed U-Net in Section 4.2 encouragingly reduces the loss from 4.86% to 1.43%. Thus the uneven color distortion is well corrected (Fig. 7(c)).

To evaluate the upper bound of the EPI spectrum loss in a dense light field, we have counted the energy loss of the spectrum over 10,000 EPIs which are reconstructed from the EFS containing all the scene depth ranges for several LFs and summarized the results in Fig. 8(a). In these EPI spectra, the maximum loss is 9.03%, the minimum loss is 0.31%, and the average loss is 1.94%. It can be seen that the EPI spectrum loss generally has a sparse ($\leq$ 5%) distribution. In addition, the loss distribution on the Disney LFs [39] is flatter and more spread out than that on the synthetic [38] and real LFs [29]. We attribute this to: 1) the background texture of the Synthetic LFs is relatively simple, while the texture of the Disney LFs is more complex; 2) the scene depth of the synthetic data is primarily concentrated within a limited interval, while the depth distribution of the Disney LFs is more divergent. Moreover, Fig. 8(b) shows the distribution of the EPI spectrum loss with the subsequent U-Net optimization. It is obvious that the energy loss is further reduced after the optimization.

Table 2
Average PSNR and SSIM of Fig. 9.

Refocus range d_{range}	0.8×	1.0×	1.2×	1.0×	1.0×
# EFSs N_f	200	200	200	104	296
PSNR↑	35.05	37.54	37.96	32.92	37.98
SSIM↑	0.865	0.952	0.959	0.823	0.961

5.3. Parameter analysis

We empirically validate the influences of the refocus range and the EFS sampling on the reconstructed EPI by performing the following parameter analysis. In these experiments, we use our synthetic LFs with 15× downsampling.

The refocus operation range. The refocus range is set to $0.8d_{range}$, d_{range} and $1.2d_{range}$, respectively. As mentioned in Section 5.2, with 15× downsampling of the original light field, d_{range} is the scene disparity range ($[15d_{\min}, 15d_{\max}]$). Fig. 9(a)–(c) show the qualitative comparisons with different refocus ranges on the tree root scene. Quantitative analysis, in terms of average PSNR and SSIM, is summarized in the 3rd and 4th rows of Table 2. As shown in Fig. 9(b), partial tree root has not been reconstructed. The scene structure can not be reconstructed completely when the refocus range is too narrow.

The EFS layer number N_f. The number of refocus layers N_f is set to 104, 200 and 296, respectively. Fig. 9(a)–(e) show the qualitative comparisons with different numbers of EFS layers. The 2nd, 5th, and 6th rows of Table 2 show the quantitative comparison on the reconstructed LFs. It is noticed that insufficient focal layers, of which the count is smaller than the minimum focal layer count N_{fmin} (see Section 3.4), *i.e.*, $N_f = 104$, will cause performance degradation (32.92/0.823 *vs* 37.54/0.952 in terms of PSNR and SSIM respectively). In this case, aliasing appears in the focal

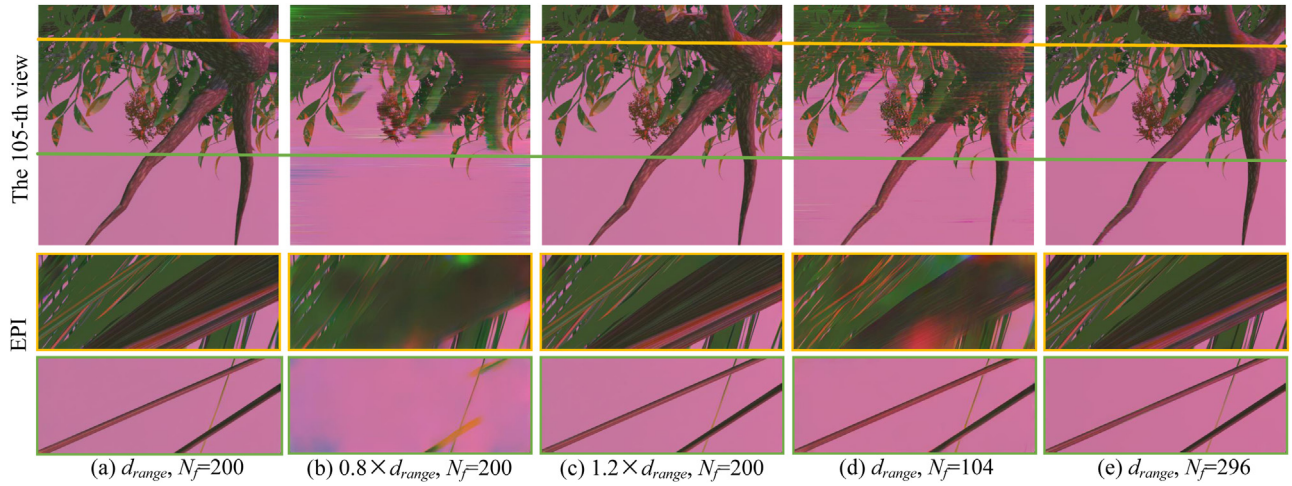


Fig. 9. Qualitative comparisons regarding the refocus operation range and the EFS layer number on a synthetic light field. The top row shows reconstructed views with different parameters. The remaining two rows show the reconstructed EPIs corresponding to the yellow and green lines in the reconstructed view respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Quantitative comparisons with SOTAs under different downsampling rates.

		Syn. LFs		Real LFs [29]		Bike	Church	Couch	Statue
		10×	15×	10×	15×	5×	10×	10×	10×
Wu et al. [13]	PSNR↑	35.48	34.77	36.14	34.92	31.13	32.64	32.01	32.63
	SSIM↑	0.851	0.813	0.877	0.825	0.708	0.721	0.724	0.698
	LPIPS↓	0.060	0.093	0.050	0.078	0.110	0.080	0.113	0.098
LLFF [37]	PSNR↑	37.29	36.46	39.74	37.02	35.51	**38.85**	37.25	38.53
	SSIM↑	0.941	0.922	**0.964**	0.925	0.875	0.962	0.916	0.956
	LPIPS↓	0.059	0.089	**0.039**	0.079	0.082	**0.051**	0.084	0.049
DILF [28]	PSNR↑	34.28	33.73	34.13	33.88	33.64	31.81	32.69	32.77
	SSIM↑	0.810	0.754	0.845	0.799	0.743	0.719	0.752	0.641
	LPIPS↓	0.085	0.106	0.087	0.093	0.096	0.087	0.098	0.090
NeLFRSE [33]	PSNR↑	37.56	36.27	**40.57**	37.26	35.45	37.66	39.18	38.17
	SSIM↑	0.949	0.904	0.941	0.917	0.915	0.920	0.905	0.938
	LPIPS↓	0.058	0.091	0.062	0.073	**0.081**	0.071	0.097	0.065
Ours	PSNR↑	**39.36**	**37.54**	40.18	**37.74**	**36.77**	37.95	**43.05**	**40.82**
	SSIM↑	**0.963**	**0.952**	0.948	**0.938**	**0.939**	**0.964**	**0.928**	**0.959**
	LPIPS↓	**0.056**	**0.088**	0.043	**0.072**	0.085	0.060	**0.079**	**0.041**

stack, which leads to over-smooth textures in Fig. 9(d). In addition, when N_f is increased from 200 to 296, only a slight improvement is reported (37.54/0.952 *vs* 37.98/0.961). Hence once the number of EFS layers N_f meets the minimum sampling rate requirement, continuously increasing EFS layers would not bring obvious improvements in the performance.

5.4. Comparisons with SOTAs

Table 3 shows the average PSNR/SSIM/LPIPS [40] measurements with different downsampling rates on both synthetic and real LFs. Qualitative comparisons between different methods on several test scenes under 15× downsampling rate are shown in Figs. 10–12 respectively.

5.4.1. Synthetic light field datasets

We evaluate the proposed method using our synthetic light field datasets under 10× and 15× downsampling rates. Qualitative results under 15× downsampling (maximum disparity up to 8px) are shown in Fig. 10.

As shown in Fig. 10(b), ghosting artifacts are visible around the boundary region in the result by Wu et al. [13], which are caused by the limited receptive field of their network. Also, the Gaussian convolution kernel is only effective for small disparity. The MPI-based LLFF [37] tends to assign high opacity to incorrect layer for the region with ambiguous/repetitive texture or moving content between input images, which will cause floating or blurred patches around the boundary region (see the boundary region of the pot in Fig. 10(c)). DILF [28] is built upon depth estimation, so an inaccurate depth map leads to errors in the edges of the reconstructed view (see the error map and the EPI of Fig. 10(d)). The NeRF-based NeLFRSE [33] is optimized pixel by pixel, so it has obvious errors at the boundary regions (see the zoom-in rectangles of Fig. 10(e)). In comparison, the proposed EFS-based reconstruction produces clear boundaries (as shown in Fig. 10(d)).

Fig. 13 shows the PSNR and SSIM measurements for each reconstructed view on the synthetic light field under 10× and 15× downsampling rates. Due to the shearing process used in our method (Eq. (1a)), the more marginal views there are, the more image information will be sheared out of the image. Thus the reconstruction results are not satisfactory on these marginal views. Still, the overall performance of our method is better than the SOTAs, especially for larger disparity scenarios. The 3rd and 4th columns of Table 3 list the quantitative measurements on synthetic LFs under different downsampling rates, which further validates the superiority of the proposed method.

5.4.2. Real LFs captured with a plenoptic camera

We also evaluate the proposed approach using the Lytro light field dataset [29], which contains massive static scenes, such as bi-

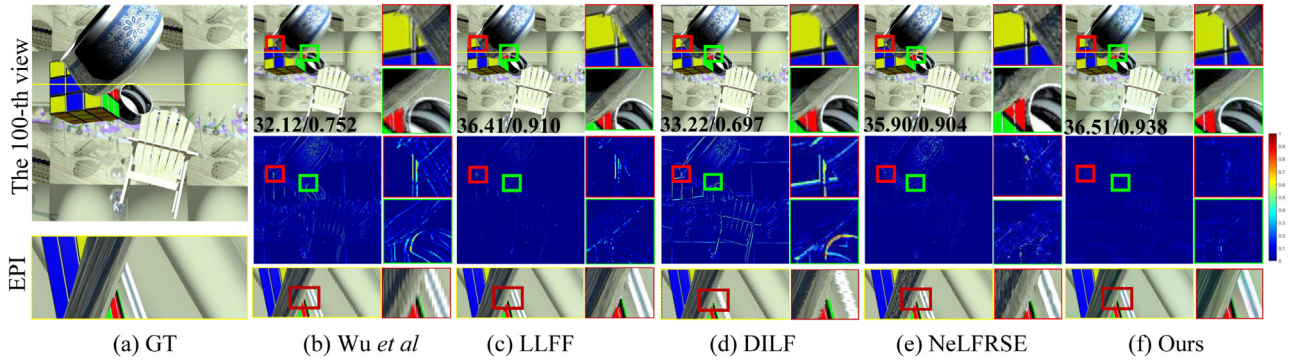


Fig. 10. Comparison on the synthetic light field dataset (15× downsampling). For each result, the reconstructed view, the error map, two close-up regions and the reconstructed EPI are provided.

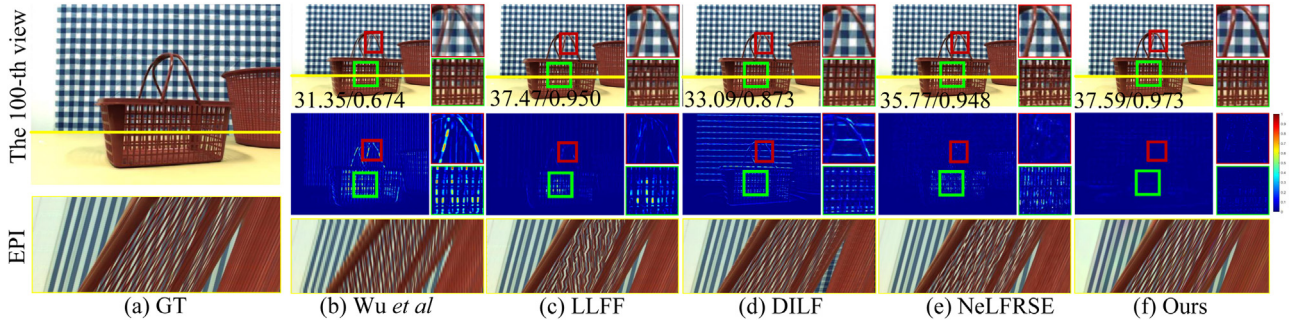


Fig. 11. Comparison on the real light field dataset (15× downsampling).

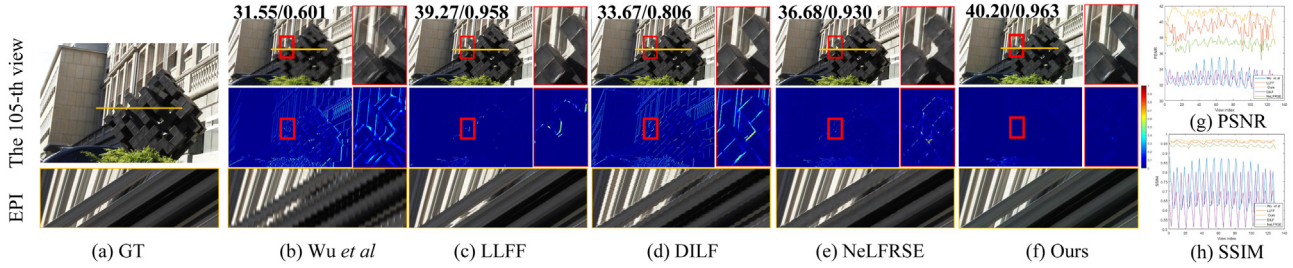


Fig. 12. Comparison on the camera array light field dataset (15× downsampling).

cycles, toys, and plants. These scenes are challenging in terms of abundant colors and complicated occlusions.

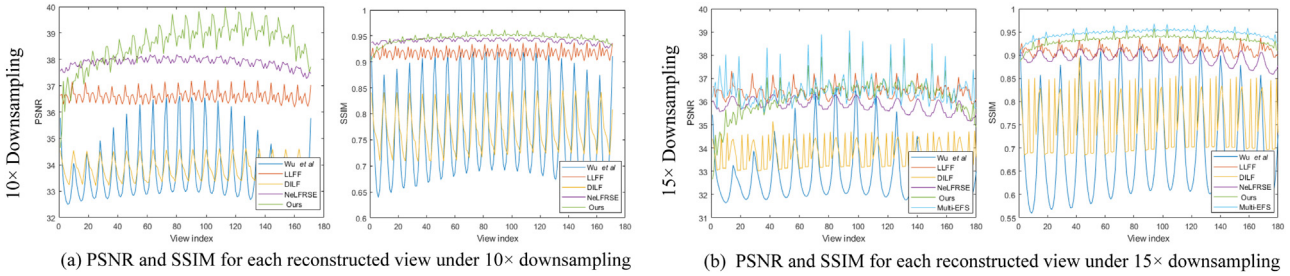
Fig. 11 shows the reconstruction results on the basket scene under 15× downsampling. There exist many thin structures in the scene, such as the basket handle. The texture on such a thin structure changes very fast, which results in difficulties for reconstruction. We can see that severe ghosting artifacts occur around the basket handle in the results by Wu et al. [13] and reconstructed views are inconsistent (Fig. 11(b)). Similarly, a fuzzy phenomenon appears in the results by LLFF [37]. DILF [28] still suffers from the edge reconstruction errors and inconsistent views. NeLFRSE [33] could not provide reliable results in heavy occlusion areas such as the basket in Fig. 11(e). By reconstructing the dense light field in the frequency domain, our method is less sensitive to spatial contents, and thus capable of producing high-quality and consistent view reconstruction.

Fig. 14 shows PSNR and SSIM measurements for each reconstructed view of Fig. 11 under 10× and 15× downsampling. The PSNR value of our method is lower than that of NeLFRSE [33] under 10× downsampling, however, at significant disparity (under 15× downsampling), our method provides better reconstruction results over 95% views. Quantitative comparisons in terms of PSNR, SSIM, and LPIPS are listed in the 5th and 6th columns of Table 3.

5.4.3. Real LFs captured with a camera array

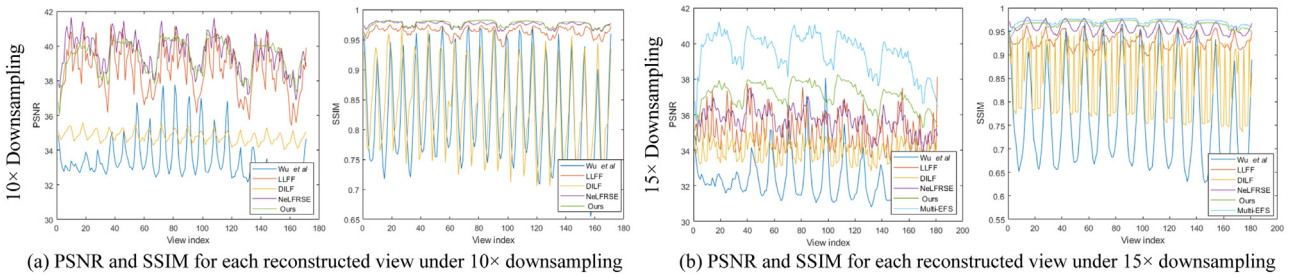
In order to verify the effectiveness of our method under wide baseline and large disparity conditions, we further evaluate the proposed approach using the Disney LFs [39] which are captured by a camera array.

Fig. 12 shows the reconstruction results on the statue scene under 10× downsampling (maximum disparity up to 15px). Due to the limited receptive field of the network, the results by Wu et al. [13] show serious aliasing effects on all the foreground objects (Fig. 12(b)). Due to the memory limitation, there is a trade-off between the image resolution and the layers of MPIs utilized by LLFF [37], which leads to a performance degradation for large disparity areas with a high-resolution input. Also severe artifacts appears in the regions with repetitive patterns, large disparity and occlusions, as shown in the zoom-in rectangles in Fig. 12(c). This is a common failure mode for the methods using texture matching cues for inferring depth. DILF [28] cannot reconstruct the view at an arbitrary position (only reconstruct the middle three views between two input views), so in the case of a large parallax, there exists serious content inconsistency across views (see the EPI of Fig. 12(d)). Since NeLFRSE [33] requires extensive input views to learn the mapping between the input rays and the RGB values, the occlusion boundaries are blurred when there are insufficient input



(a) PSNR and SSIM for each reconstructed view under 10× downsampling (b) PSNR and SSIM for each reconstructed view under 15× downsampling

Fig. 13. PSNR and SSIM of the reconstructed views of Fig. 10.



(a) PSNR and SSIM for each reconstructed view under 10× downsampling (b) PSNR and SSIM for each reconstructed view under 15× downsampling

Fig. 14. PSNR and SSIM of the reconstructed views of Fig. 11.

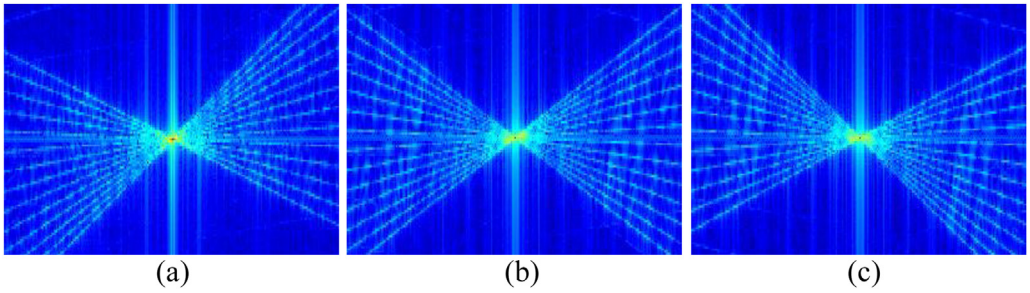


Fig. 15. EFSs by different reference views under 15× downsampling (the original light field has 200 views). From (a) to (c), the index of the reference views is 71, 101 and 131, respectively.

views. In contrast, thanks to the depth-independent characteristic, our proposed method shows better performance for large disparities. Furthermore, it is noticed that the proposed method maintains better view consistency compared with other methods. Quantitative results on several scenes of the Disney LFs are shown in the last four columns of Table 3.

5.5. Multi-reference-view results

As shown in Figs. 13 and 14, the reconstruction performance for some marginal views by our method (only one center reference) has been slightly below than that of the SOTAs. This is because the center reference view does not contain enough side-view information when the baseline is large. In this section, we use multi-reference views to build multi-EFSs to provide more information for marginal view reconstruction.

Fig. 15 shows reconstructed multi-EFSs, which are built using the 30th view ahead of the center view ($u_{ref} = u_{cen} - 30$), the center reference ($u_{ref} = u_{cen}$) and the 30th view behind the center view($u_{ref} = u_{cen} + 30$). Fig. 13(b) and 14(b) show PSNR and SSIM measurements for each reconstructed view of Figs. 10 and 11. It can be seen that utilizing multiple EFSs significantly improves the overall performance of our EFS-based method, especially for the edge views.

5.6. Full parallax results

To verify our method's capability to reconstruct the 4D light field with both horizontal and vertical views, we provide the reconstruction of the full parallax light field in this section. The sequential solution for full parallax is reconstructing the views containing vertical disparities after all the views containing horizontal disparities have been reconstructed, which is illustrated using an array of 17×17 full parallax reconstructed from the 9×9 input views (marked in black) in Fig. 16. As shown in Fig. 16(b), the views marked in green are firstly reconstructed in the horizontal parallax reconstruction step, and the views marked in blue are later reconstructed in the vertical parallax reconstruction step. We validate this solution on the Lego dataset [41] and show the qualitative and quantitative results in Fig. 17.

Due to the larger disparity ($[-9, 7]$) of the Lego dataset, there exist apparent artifacts near boundary regions and discontinuous EPIs in both the results by Wu et al. [13] and LLFF [37]. Additionally, LLFF [37] usually requires a large dataset for model training, while our method is capable of learning the relation between views and spectrum lines in the frequency domain from a relatively small training dataset. NeLFRSE [33] achieves good results benefiting from a large number of input views (81) in this set, however, it requires additional time to retrain the network for dif-

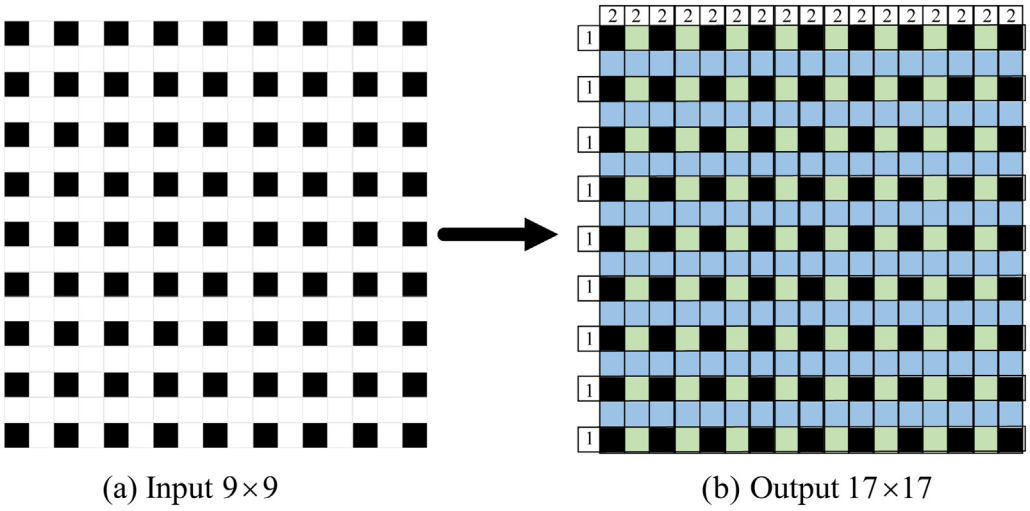


Fig. 16. Illustration of the sequential solution for full parallax light field reconstruction. (a) 9 × 9 input views. (b) 17 × 17 output views.

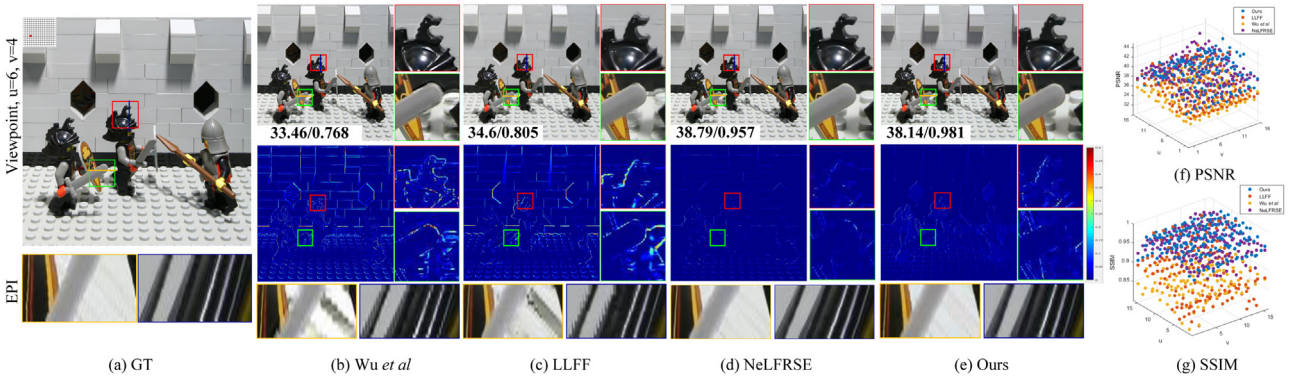


Fig. 17. Comparison on the Lego dataset (from 9 × 9 to 17 × 17 views).

Table 4
The average PSNR and SSIM on the Lego dataset.

	Wu et al. [13]	LLFF [37]	NeLFRSE [33]	Ours
PSNR↑	36.29	37.71	**39.25**	39.02
SSIM↑	0.862	0.895	0.954	**0.965**

ferent scenes. Moreover, its lower SSIM value tells that the pixel-by-pixel optimization cannot keep the image structure well. The experimental results show that our proposed method can generate clear edges and preserve cross-view consistency. Table 4 shows the average PSNR and SSIM on the Lego dataset.

5.7. Limitation

At present, the energy loss of the spectrum still exists, which may lead to uneven color distortions, as shown in the red rectangle of Fig. 7. A possible solution could be to adapt the spatio-frequency combined method, or adding a visual channel to compensate for the energy loss in the frequency domain. In addition, although the proposed method outperforms the SOTA methods on both view reconstruction quality and cross-view consistency preservation, the shearing operation may cause a moderate decline in the reconstruction performance for the marginal views (Section 5.4). This could be addressed by introducing more reference views during the construction of the focal stack to provide more scene information.

6. Conclusions

In this paper, we have extended the focal stack spectrum theory and presented the EFS for representing the light field reversibly, providing the theoretical basis for dense light field reconstruction from a sparse one using the EFS. We analyse the EFS sampling problem, and derive the analytical function of the minimal focal stack layers according to the scene distribution function, camera parameters, the number of views, and scene depth. In the implementation, we first reconstruct the EFS of a dense light field from a sparse one using a dual-stream network. Then this dense EFS is projected to the EPI, which is finally fed into a subsequent U-Net to eliminate the uneven color distortion. Experimental results show that the proposed method exhibits superior performance under many challenging conditions, such as large disparities and complex occlusions.

The proposed method is capable of preserving the cross-view consistency, however, due to the slight energy leakage in the reconstructed spectrum, there exists an uneven color distortion. In the future, we will try to minimize the energy loss and address the issue in a learnable Fourier network.

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.

Acknowledgements

This work was supported by NSFC under Grant 62031023, Grant 61801396 and Grant 62101242.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2023.109551.

References

[1] E.H. Adelson, J.R. Bergen, The plenoptic function and the elements of early vision, Comput. Models Vis. Process. 1 (2) (1991) 3–20.
[2] M. Levoy, P. Hanrahan, Light field rendering, in: ACM SIGGRAPH, ACM, 1996, pp. 31–42.
[3] S.J. Gortler, R. Grzeszczuk, R. Szeliski, M.F. Cohen, The lumigraph, in: ACM SIGGRAPH, 1996, pp. 43–54.
[4] K. Gedamu, Y. Ji, Y. Yang, L. Gao, H.T. Shen, Arbitrary-view human action recognition via novel-view action generation, Pattern Recognit. 118 (2021) 108043.
[5] C. Zhang, G. Hou, Z. Zhang, Z. Sun, T. Tan, Efficient auto-refocusing for light field camera, Pattern Recognit. 81 (2018) 176–189.
[6] T. Georgiev, K.C. Zheng, B. Curless, D. Salesin, S.K. Nayar, C. Intwala, Spatio-angular resolution tradeoffs in integral photography, Rendering Tech. 2006 (263–272) (2006) 21.
[7] S. Wanner, B. Goldluecke, Variational light field analysis for disparity estimation and super-resolution, IEEE TPAMI 36 (3) (2014) 606–619.
[8] N.K. Kalantari, T.-C. Wang, R. Ramamoorthi, Learning-based view synthesis for light field cameras, ACM TOG 35 (6) (2016) 193:1–193:10.
[9] P.P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, R. Ng, Learning to synthesize a 4D RGBD light field from a single image, in: IEEE ICCV, 2017, pp. 2262–2270.
[10] J. Liu, N. Song, Z. Xia, B. Liu, J. Pan, A. Ghaffar, J. Ren, M. Yang, A dense light field reconstruction algorithm for four-dimensional optical flow constraint equation, Pattern Recognit. 134 (2023) 109101, doi:10.1016/j.patcog.2022.109101.
[11] B. Mildenhall, P.P. Srinivasan, R. Ortiz-Cayon, N.K. Kalantari, R. Ramamoorthi, R. Ng, A. Kar, Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, ACM TOG 38 (4) (2019) 1–14.
[12] J.-X. Chai, X. Tong, S.-C. Chan, H.-Y. Shum, Plenoptic sampling, in: ACM SIGGRAPH, 2000, pp. 307–318.
[13] G. Wu, Y. Liu, L. Fang, Q. Dai, T. Chai, Light field reconstruction using convolutional network on EPI and extended applications, IEEE TPAMI 41 (7) (2019) 1681–1694.
[14] L. Shi, H. Hassanieh, A. Davis, D. Katabi, F. Durand, Light field reconstruction using sparsity in the continuous fourier domain, ACM TOG 34 (1) (2014) 12:1–12:13.
[15] S. Vagharshakyan, R. Bregovic, A. Gotchev, Light field reconstruction using shearlet transform, IEEE TPAMI 40 (1) (2018) 133–147.
[16] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, I. So Kweon, Learning a deep convolutional network for light-field image super-resolution, in: IEEE ICCV Workshops, 2015, pp. 24–32.
[17] Y. Li, X. Wang, H. Zhu, G. Zhou, Q. Wang, Deep anti-aliasing of whole focal stack using slice spectrum, IEEE TCI 7 (2021) 1328–1340.
[18] S. Inamori, S. Yamauchi, A method of noise reduction on image processing, IEEE Trans. Consum. Electron. 39 (4) (1993) 801–805.
[19] R. Ng, Fourier slice photography, ACM Trans. Graph. 24 (3) (2005) 735744.
[20] D.G. Dansereau, O. Pizarro, S.B. Williams, Linear volumetric focus for light field cameras, ACM TOG 34 (2) (2015) 15.
[21] M. Le Pendu, C. Guillemot, A. Smolic, A fourier disparity layer representation for light fields, IEEE TIP 28 (11) (2019) 5740–5753.
[22] A. Chang, T. Sung, K. Shih, H.H. Chen, Anti-aliasing for light field rendering, in: IEEE ICME, 2014, pp. 1–6, doi:10.1109/ICME.2014.6890175.
[23] Z. Xiao, Q. Wang, G. Zhou, J. Yu, Aliasing detection and reduction in plenoptic imaging, in: IEEE CVPR, 2014, pp. 3326–3333.
[24] A. Isaksen, L. McMillan, S.J. Gortler, Dynamically reparameterized light fields, in: ACM SIGGRAPH, 2000, pp. 297–306.
[25] A. Levin, F. Durand, Linear view synthesis using a dimensionality gap light field prior, in: IEEE CVPR, 2010, pp. 1831–1838.
[26] A. Lumsdaine, T. Georgiev, et al., Full resolution lightfield rendering, Indiana Univ. Adobe Syst. Tech. Rep. 91 (2008) 92.
[27] P.P. Srinivasan, R. Tucker, J.T. Barron, R. Ramamoorthi, R. Ng, N. Snavely, Pushing the boundaries of view extrapolation with multiplane images, in: IEEE CVPR, 2019, pp. 175–184.
[28] M. Guo, J. Jin, H. Liu, J. Hou, Learning dynamic interpolation for extremely sparse light fields with wide baselines, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2450–2459.
[29] M. Guo, H. Zhu, G. Zhou, Q. Wang, Dense light field reconstruction from sparse sampling using residual network, in: Springer ACCV, 2018, pp. 50–65.
[30] H. Zhu, M. Guo, H. Li, Q. Wang, A. Robles-Kelly, Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution, IEEE TVCG 27 (6) (2021) 3019–3033.
[31] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, et al., State of the art on neural rendering, in: Computer Graphics Forum, Vol. 39, Wiley Online Library, 2020, pp. 701–727.
[32] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, NeRF: representing scenes as neural radiance fields for view synthesis, in: European Conference on Computer Vision, Springer, 2020, pp. 405–421.
[33] B. Attal, J.-B. Huang, M. Zollhöfer, J. Kopf, C. Kim, Learning neural light fields with ray-space embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19819–19829.
[34] R.C. Gonzales, R.E. Woods, Digital image processing, 2nd Edition, Prentice Hall, 2002.
[35] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE TIP 13 (4) (2004) 600–612.
[36] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Springer ECCV, 2016, pp. 694–711.
[37] B. Mildenhall, P.P. Srinivasan, R. Ortiz-Cayon, N.K. Kalantari, R. Ramamoorthi, R. Ng, A. Kar, Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, ACM TOG 38 (4) (2019) 1–14.
[38] NPU-CVPG, Dense light field datasets, (http://www.npu-cvpg.org/DenseLightField).
[39] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, M.H. Gross, Scene reconstruction from high spatio-angular resolution light fields, ACM TOG 32 (4) (2013) 73:1–73:12.
[40] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: IEEE CVPR, 2018, pp. 586–595.
[41] The new stanford light field archive, (http://lightfield.stanford.edu/lfs.html).

Yaning Li is now a PhD candidate with the School of Computer Science, Northwestern Polytechnical University. She received the MS degree from the School of Computer Science and Engineering, Xi'an University of Technology, in 2018. Her research interests include computational photography, light field computing theory and applications.

Dr. Xue Wang is an Associate Research Fellow with the School of Computer Science, Northwestern Polytechnical University. She received the BS and PhD degrees from Northwestern Polytechnical University, in 2007 and 2017, respectively. From 2012 to 2014, she studied in University of Pennsylvania as a visiting PhD student financed by China Scholarship Council. Her research interests include computer vision, computational photography and machine learning.

Dr. Hao Zhu is an Associate Researcher in the School of Electronic Science and Engineering, Nanjing University. He received the BS and PhD degrees from Northwestern Polytechnical University in 2014 and 2020, respectively. He was a visiting scholar at the Australian National University. His research interests include computational photography and optimization for inverse problems.

Dr. Guoqing Zhou is an Associate Professor in the School of Computer Science, Northwestern Polytechnical University. He received the BS degree in Computer Science from the School of Computer Science, Northwestern Polytechnical University, in 2003. He then joined Northwestern Polytechnical University as a Lecturer. In 2009 and 2013 he obtained Master and Ph.D. degrees in the School of Computer Science, Northwestern Polytechnical University. He worked as a visiting scholar in Center for Imaging Science (CIS) of The Johns Hopkins University, in 2016.9–2017.9. His research interests include computer vision and computational photography, such as 3D reconstruction, global optimization, light field imaging and processing.

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 School of Computer Science, 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, light field imaging and processing, novel view synthesis.