Sheared Epipolar Focus Spectrum for Dense Light Field Reconstruction

Yaning Li, Xue Wang, Guoqing Zhou, Hao Zhu and Qing Wang, *Senior Member, IEEE*

Abstract—This paper presents a novel technique for the dense reconstruction of light fields (LFs) from sparse input views. Our approach leverages the Epipolar Focus Spectrum (EFS) representation, which models the LF in the transformed spatial-focus domain, avoiding the dependence on the scene depth and providing a high-quality basis for dense LF reconstruction. Previous EFS-based LF reconstruction methods learn the cross-view, occlusion, depth and shearing terms simultaneously, which makes the training difficult due to stability and convergence problems and further results in limited reconstruction performance for challenging scenarios. To address this issue, we conduct a theoretical study on the transformation between the EFSs derived from one LF with sparse and dense angular samplings, and propose that a dense EFS can be decomposed into a linear combination of the EFS of the sparse input, the sheared EFS, and a high-order occlusion term explicitly. The devised learning-based framework with the input of the under-sampled EFS and its sheared version provides high-quality reconstruction results, especially in large disparity areas. Comprehensive experimental evaluations show that our approach outperforms state-of-the-art methods, especially achieves at most > 4 dB advantages in reconstructing scenes containing thin structures.

Index Terms—Epipolar Focus Spectrum (EFS), Sheared EFS, Focal stack, LF reconstruction.

1 Introduction

LIGHT field (LF) imaging has emerged as a powerful tool for capturing images with novel and unique effects, including free-viewpoint imaging [1], [2], all-in-focus imaging [3], and 4D editing [4], [5]. Due to its remarkable capabilities, LF imaging has gained considerable attention in the computational photography community [6]. Nevertheless, a fundamental challenge in LF imaging is the 'spatio-angular resolution tradeoff' [7], [8], [9], which limits the acquisition resolution of LF systems. Specifically, an increase in resolution in one dimension results in a decrease in resolution in another dimension. Hence, it is of utmost importance to explore strategies for reconstructing a densely-sampled LF from a sparsely-sampled acquisition.

In recent years, several approaches have been proposed in the literature for reconstructing high-angular resolution light fields (LFs). Despite the significant progress, several challenges remain. In the spatial domain, prevailing methods, such as those in [10], [11], [12], [13], [14], [15], [16], entail estimating depth, warping input views, and refining the textures of novel views. Despite considerable research efforts to enhance depth accuracy [10], scene representation [17], [18], and texture consistency [19], [20], it is still challenging to preserve the view-consistency of the reconstructed LF due to the inherent nature of independent view optimization employed in these methods. Moreover, occlusion accumulation in regions with large disparities causes significant artifacts. In the frequency domain, several existing works, such as those in [21], [22], [23], [24], [25], focus on Fourier spectral completion using the dimensional gap between the 3D focal stack and the 4D LF [22], sparsity in the continuous Fourier domain [23], or band-consistency in the shearlet domain [24]. However, these methods only model the Fourier features of non-overlapped and continuous epipolar plane image (EPI) lines, and are incapable of modeling EPI lines with intersection and discrete EPI points/segments, i.e., occlusion and large disparities. The fundamental incompleteness of these models results in artifacts in the reconstructed occlusion boundaries.

A novel representation, known as the Epipolar Focus Spectrum (EFS), has been recently introduced for the purpose of 4D LF anti-aliasing refocusing [26] and dense reconstruction [27]. The EFS possesses semantic attributes that are invariant to scene depth. Rendering a complete EFS literally contributes to the task of view-consistent LF reconstruction. However, the existing literature [26], [27] is predominantly focused on the distribution of the magnitude spectrum within the EFS, with limited analysis of the transformation between EFSs derived from the sparse and dense LF data.

In this paper, we present an enhanced formulation of the EFS theory and promote its applicability in the domain of LF reconstruction. Our analysis reveals that the EFS of a densely-sampled LF can be represented as a linear combination of three constituent terms, namely, the EFS of the corresponding sparsely-sampled LF, a sheared sparse EFS, and a cross term. Based on this insight, we introduce a novel shearing module that facilitates the preprocessing of the input sparse EFS to generate a coarse approximation of the dense EFS. Subsequently, we propose an Occlusion-aware Dual-stream U-Net (ODU-Net) architecture that is specifically designed to refine the coarse approximation. Finally, the reconstructed LF is obtained via the application of an inverse Fourier slice photography technique [28] to

- *Y. Li, X. Wang, G. Zhou and Q. Wang (corresponding author) are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (E-mail: qwang@nwpu.edu.cn).*
- *H. Zhu is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: zhuhao_photo@nju.edu.cn).*

- *The work was supported by NSFC under Grants 62031023 and 62101242.*

Manuscript received April 19, 2005; revised August 26, 2015.

the refined EFS. Compared to previous methods, the shear operation employed in our approach reduces the difficulty of spectrum completion during the network training, resulting in improved accuracy. We demonstrate the efficacy of our proposed approach through extensive experimental evaluations conducted on both synthetic and real-world LF datasets (as detailed in Sec. 5).

The paper makes several significant contributions, including:

1) The formal breakdown of the EFS based on a sparsely sampled EFS. The relationship between the EFSs of sparse and dense LFs is analyzed and concluded, providing a solid theoretical foundation for high-quality, view-consistent LF reconstruction.

2) A specially designed ODU-Net with phase modulation for dense LF reconstruction. When applied to the sheared sparse EFS, ODU-Net not only alleviates the training difficulty of spectrum completion, but also improve the accuracy of view reconstruction.

2 Related Work

2.1 LF representation and reconstruction in the image domain

Levoy and Hanrahan [29] introduce a 4D LF representation, denoted by $L(u, v, x, y)$, which uses a two-parallel-plane model. Here, (u, v) and (x, y) represent the intersections of the ray with angular/camera and spatial/image planes, respectively. Various LF reconstruction methods have been proposed. One common approach involves estimating the scene depth [30], [31], warping input views to synthesize novel views [11], [19], and gradually applying convolutional neural networks (CNNs) for LF reconstruction. Taking the sparse 4D LF as the network input, Wang et al. [32] propose a pseudo-4D neural network that directly reconstructs the LF in the angular domain. Yeung et al. [33] propose an LF reconstruction method that alternates convolution of the angular and spatial dimensions, thereby improving the efficiency of viewpoint reconstruction. Kalantari et al. [14] estimate the depth and color of each viewpoint sequentially using two convolutional neural networks. Yoon et al. [34] achieve super-resolution reconstruction of adjacent views using a data-driven learning method.

Based on the structural properties of angular-dimension sampling in the light field, the EPI can be utilized to represent the 4D LF [35]. The use of EPI for view reconstruction enables better maintenance of the angular consistency of the LF [36]. As a result, EPI has gained widespread use in angular super-resolution. Wu et al. [37] convert the angular domain reconstruction of a 4D LF into a one-dimensional super-resolution of the 2D EPI. Guo et al. [38] transform the LF reconstruction problem into the prediction of the residual between the EPIs of dense and sparse LF. Zhu et al. [36] further improve the super-resolution performance on EPI in large disparity areas by introducing an LSTM module. Vagharshakyan et al. [24] employ an iterative regularized reconstruction by utilizing a sparse representation of the underlying EPIs in the shearlet domain.

The 4D LF representation exhibits high redundancy, and the EPI-based technique fails to provide accurate texture information for the scene. In contrast, refocused images can describe the scene's texture information and convey the relative depth of objects through blurring [35]. Consequently, the focal stack is commonly used for LF representation [22], and is widely adopted for LF reconstruction. Levin and Durand [22] propose a 3D focal stack spectral completion method for dense LF reconstruction. Sakamoto et al. [39] encode the focal stack with a wavelet-based method, and then reconstruct the LF from the focal stack using a linear view synthesis method [22].

Superpixel [4], geometry-aware graph [40], [41], hypergraph [42] and multiplane images (MPIs) [17], [43] can also represent LFs considering the depth and texture information. Superpixel and hypergraph representations are generally used for LF segmentation [4], [5], [42] and compression [40], [44], [45]. Srinivasan et al. [43] propose to utilize the MPI representation to synthesize the viewpoint from a narrow baseline stereo pair. Based on a learning gradient descent algorithm, Flynn et al. [46] generate MPIs from a sparse set of viewpoints for LF reconstruction. Mildenhall et al. [18] further improve the performance of LF reconstruction by fusing multiple MPIs. Tucker et al. [47] utilize a single image to construct MPIs and perform view synthesis.

These LF representations are highly correlated with scene depth or contain redundant information. Features extracted or learned from LF with a small disparity range may lead to artifacts when applied to LF with a large disparity range.

2.2 LF representation and reconstruction in the Fourier domain

Ng [28] proposes the Fourier slice model for LF based on the Fourier slicing theory introduced by Levoy [48]. This model provides a 2D Fourier slice representation of the LF. In subsequent studies, Shi et al. [23] and Levin and Durand [22] achieve LF reconstruction for small samples by analyzing the sparsity of Fourier slices, respectively.

Le Pendu et al. [25] introduce the Fourier disparity layer (FDL) as a novel representation and utilize it for LF reconstruction. The FDL representation enables the generation of various views through a straightforward process of shifting and filtering.

Based on disparity cues and Fourier slicing theory, Li et al. [26] propose an EFS representation for LF. They find that each spectral line in the EFS corresponds to a viewpoint of the LF. Therefore, the dense LF reconstruction task is formulated as an EFS completion problem and solved using a learning framework [27].

The Fourier domain LF representations described earlier do not account for the occlusion present within a scene, leading to a lack of sparse spectrum features. In this study, our objective is to examine the correlation between the EFS derived from the sparsely sampled LF and the EFS associated with the targeted reconstructed views. To accomplish this, we introduce an occlusion model that minimizes the error in LF reconstruction.

3 Theoretical Analysis

To better analyze the relationship between the sparse EFS and dense EFS, we will first introduce the background and

TABLE 1
Notations of symbols

Notation	Definition
$L(u,v,x,y)$	4D LF
$E_n(u,x)$	EPI with n views
$E_{kn}(u,x)$	EPI with $k*n$ views
$\mathcal{E}_n(\omega_u,\omega_x)$	Fourier spectrum of $E_n(u,x)$
$\mathcal{F}_n(f,\omega_x)$	Rearranged $\mathcal{E}_n(\omega_u,\omega_x)$
$F_n(f,x)$	Focal stack formed from $E_n(u,x)$
f	Relative focused depth in building $F_n(f,x)$
$\Delta\alpha$	Refocus step for building $F_n(f,x)$
u_{ref}	Reference view for building $F_n(f,x)$
$\Delta F(f,x)$	Focal stack from additional $(k-1)n$ views
$EFS_n(\omega_f,\omega_x)$	Epipolar focus spectrum of $E_n(u,x)$
$\Delta EFS(\omega_f,\omega_x)$	Fourier spectrum of $\Delta F(f,x)$
$FT_{1D}(\cdot)$	1D Fourier transform operator
$FT_{2D}(\cdot)$	2D Fourier transform operator

the notations used throughout the paper. Then, a complete theoretical analysis of the sheared EFS is derived.

3.1 Notations

Given a 4D LF $L(u,v,x,y)$, $E_n(u,x)$ is the EPI by fixing $(v,y)=(v^*,y^*)$, where n is the number of views in the EPI. $E_{kn}(u,x)$ refers to the EPI with k-times dense sampling. Note that the same unit is employed for modeling both $E_n(u,x)$ and $E_{kn}(u,x)$, i.e., the set of view indices in $E_n(u,x)$ is $1,k+1,2k+1,\ldots,(n-1)k+1$, while the set of view indices in $E_{kn}(u,x)$ is $1,2,\ldots,nk$. $\mathcal{E}n(\omega_u,\omega_x)$ represents the Fourier spectrum of $En(u,x)$, and $\mathcal{F}n(f,\omega_x)$ is the rearrangement of $\mathcal{E}n(\omega_u,\omega_x)$ after slicing.

$F_n(f,x)$ is the focal stack formed from $E_n(u,x)$ where f refers to the relative focused depth. It should be noted that $F_n(f,x)$ is constructed with a refocus step of $\Delta\alpha$. The reference view in $F_n(f,x)$ is u_{ref}, which is set as the central view throughout the paper. $\Delta F(f,x)$ refers to the focal stack formed from additional $(k-1)n$ views, *i.e.*, the views $\{2,3,...,k,k+2,k+3,...,2k,...,(n-1)k+2,...,nk\}$. $EFS_n(\omega_f,\omega_x)$ is the epipolar focus spectrum of $E_n(u,x)$, or in other words, the Fourier spectrum of $F_n(f,x)$. $\Delta EFS(\omega_f,\omega_x)$ is the Fourier spectrum of $\Delta F(f,x)$.

$FT_{1D}(\cdot)$ and $FT_{2D}(\cdot)$ are 1D and 2D Fourier transforms to the signal $\cdot$, respectively. Tab. 1 lists all symbols used throughout the paper.

3.2 Background

Given a 2D EPI $E_n(u,x)$, its EFS could be constructed in two ways [27]. In the spatial domain, the focal stack $F_n(f,x)$ is firstly constructed and then processed by the 2D Fourier transform,

$$F_n(f,x)=\frac{1}{n}\sum_{u=1}^{n}E_n(u,x+f(u-u_{ref})), \tag{1a}$$

$$EFS_n(\omega_f,\omega_x)=FT_{2D}(F_n(f,x)). \tag{1b}$$

In the Fourier domain, the spectrum $\mathcal{E}_n(\omega_u,\omega_x)$ is firstly obtained and then rearranged in a slice way. Finally, the 1D Fourier transform is applied to the rearranged EPI spectrum,

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

$$EFS_n(\omega_f,\omega_x)=FT_{1D}(\mathcal{F}_n(f,\omega_x)). \tag{2b}$$

Please refer to the original EFS paper [27] and the Fourier slice photography theory for the equivalence of the above equations.

According to [26], [27], the pattern of EFS is independent/irrelevant to the scene depth, and there is a one-to-one correspondence between each EFS line and the view in the LF. All lines pass through the origin. The slope of each line is determined by the refocus step $\Delta\alpha$ and the interval between the current view and the reference view.

3.3 Representing a dense focal stack from a sparse one

According to the one-to-one correspondence between the view and EFS line, the task of reconstructing a dense LF E_{kn} from a sparse one E_n could be modeled as 'inserting' the energy bands of other $(k-1)n$ views into EFS_n and then modulating the phase. It is the key to the 'inserting' process by representing the information of other $(k-1)n$ views from the input n views.

Let us start with the focal stack representation. According to the linear-weighting composition nature of the focal stack generation from EPI, the focal stack F_{kn} could be decomposed as the sum of F_n and ΔF,

$$F_{kn}=\frac{1}{k}F_n+\frac{k-1}{k}\Delta F. \tag{3}$$

Correspondingly, Eq. 3 could be rewritten in the Fourier domain according to the linear property of Fourier transform,

$$EFS_{kn}=\frac{1}{k}EFS_n+\frac{k-1}{k}\Delta EFS. \tag{4}$$

In the following paragraphs, we will successively analyze the expansion of Eqs. 3 and 4 in the single-point and multi-point scenes.

3.3.1 Scene with single-point

To simplify the derivation of ΔF from F_n, we first focus on a simple scene in which only one point P exists in the space. **Spatial domain.** Fig. 1 gives an intuitive illustration of the case of one point with $k=2$ and $n=3$ sampling. Fig. 1(a) shows the focal stack F_n. The pixel p lies at the focused depth f_p^*. p_1, p_3 and p_5 come from the views u_1, u_3 and u_5 respectively and lie at a defocus depth f. According to the Lambertian assumption, the intensities of p_1, p_3, and p_5 could be formulated as a mixture of the intensity of p and the background. Considering only one point exists in the space, *i.e.*, the intensity of the background is a constant value C. As a result, the intensity of $p_j(f,x_j)$ could be modeled as

$$\forall j\in\{1,3,5\},F_n(f,x_j)=\frac{1}{3}F_n(f_p^*,x)+\frac{2}{3}C. \tag{5}$$

Fig. 1(b) shows the focal stack ΔF formed from other 3 views and we also have

$$\forall j\in\{2,4,6\},\Delta F(f,x_j)=\frac{1}{3}\Delta F(f_p^*,x)+\frac{2}{3}C. \tag{6}$$

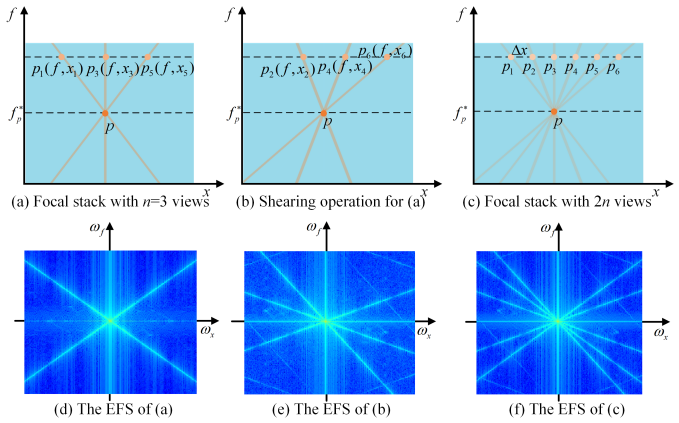


Fig. 1. Analysis of the shear operation on the focal stack and EFS. (a) $F_n(f,x)$ comprised of $n=3$ views. (b) $\Delta F(f,x)$ shearing from (a). (c) $F_{kn}(f,x)$ comprised of $2*n=6$ views. (d) EFS of (a). (e) EFS of (b), which can be also obtained by the shearing and linear phase modulation from (d). (f) EFS of (c).

Under the Lambertian assumption, we have $F_n(f_p^*,x)=\Delta F(f_p^*,x)$. By combining Eqs. 5 and 6, the following equation holds,

$$\forall i\in\{1,3,5\}, j\in\{2,4,6\}\,, F_n(f,x_i)=\Delta F(f,x_j). \tag{7}$$

Since the baseline between neighboring views in F_n is equal to the baseline in ΔF (for any $i\in\{2,4,6\}$, $u_i-u_{i-1}=u_{i+1}-u_i$), ΔF could be represented by shearing F_n,

$$\Delta F(f,x)=F_n(f,x+\Delta x), \tag{8}$$

where Δx models the interval between the lines pp_1 and pp_2 along the focus line f in Fig. 1(c) (*e.g.*, $\Delta x=0$ when scanning the line f_p^*). According to the radius of defocus blur in focal stack construction [26], we have $\Delta x=\Delta u(f_p^*-f)$, where $\Delta u=u_2-u_1$ describes the distance between the first set of views $\{u_1,u_3,u_5\}$ and the second set of views $\{u_2,u_4,u_6\}$. As a result, Eq. 8 can be expanded as

$$\Delta F(f,x)=F_n(f,x+\Delta u(f_p^*-f)). \tag{9}$$

Furthermore, by extending the simple example ($k=2$, $n=3$) to a general case, Eq. 9 is reformulated as

$$\Delta F(f,x)=\frac{1}{k-1}\sum_{j=1}^{k-1}F_n(f,x+\Delta u_j(f_p^*-f)), \tag{10}$$

where $\Delta u_j=u_{1+j}-u_1$, referring to the distance between the original set of views $\{u_1,u_{k+1},u_{2k+1},...,u_{nk+1}\}$ and the j-th set of reconstructed views $\{u_{1+j},u_{k+1+j},...,u_{nk+1+j}\}$. Consequently, Eq. 3 could be rewritten as

$$\begin{aligned}F_{kn}(f,x)=&\frac{1}{k}F_n(f,x)+\frac{k-1}{k}\Delta F(f,x)\\ =&\frac{1}{k}\left(F_n(f,x)+\sum_{j=1}^{k-1}F_n(f,x+\Delta u_j(f_p^*-f))\right).\end{aligned} \tag{11}$$

Fourier domain. The shearing process described in ΔF of Eq. 11 can be formulated in an affine transformation

$$\begin{pmatrix}f'\\x'\end{pmatrix}=\underbrace{\begin{pmatrix}1&0\\-\Delta u_j&1\end{pmatrix}}_{\boldsymbol{A}}\begin{pmatrix}f\\x\end{pmatrix}+\underbrace{\begin{pmatrix}0\\f_p^*\Delta u_j\end{pmatrix}}_{\boldsymbol{x}_0^p}. \tag{12}$$

According to the affine theorem of Fourier transform [49], the affine transform in the spatial domain is equivalent to a combination of affine transform, amplitude scaling, and phase modulation in the Fourier domain. In short, $\boldsymbol{\omega}=(\omega_f,\omega_x)^\top$, therefore,

$$\begin{aligned}&EFS_{kn}\\ =&\frac{1}{k}EFS_n+\frac{k-1}{k}\Delta EFS\\ =&\frac{1}{k}\left(EFS_n+\sum_{j=1}^{k-1}\frac{1}{|\det(\boldsymbol{A})|}e^{2\pi i\boldsymbol{x}_0^{p\top}\boldsymbol{A}^{-\top}\boldsymbol{\omega}}EFS_n\left(\boldsymbol{A}^{-\top}\boldsymbol{\omega}\right)\right)\\ =&\frac{1}{k}\left(EFS_n+\sum_{j=1}^{k-1}e^{2\pi i\boldsymbol{x}_0^{p\top}\boldsymbol{A}^{-\top}\boldsymbol{\omega}}EFS_n\left(\boldsymbol{A}^{-\top}\boldsymbol{\omega}\right)\right)\end{aligned} \tag{13}$$

where i represents an imaginary number satisfying $i^2=-1$.

Comparing Eqs. 11 and 13, it is found that the shearing process can only be achieved in the spatial domain when the focused depth of the point is known, however, the shearing process is separable in the Fourier domain, and the shearing in the power spectrum is independent of the focused depth of the point.

3.3.2 *From single-point to multi-point scene*

The above analysis focuses on the case with only one point. When there exist many points in the scene, Eq. 11 is inappropriate since the focused depth varies from point to point. To analyse the case with multiple points, additional masks $\{M^p\}_{p=1}^{N_p}$ are introduced to separate F_n into multiple focal stacks $\{F_n^p\}_{p=1}^{N_p}$ where each F_n^p is a focal stack including only one point and N_p denotes the number of points in the scene. Apart from this, the cross-view phenomenon appears (see the blue box in Fig. 2). It is essential to add the third term F_n^{cross} to model the cross-view effects. As a result, Eq. 11 could be modified as

$$\begin{aligned}&F_{kn}(f,x)\\ =&\frac{1}{k}\left(F_n(f,x)+\sum_{j=1}^{k-1}\sum_{p=1}^{N_p}M^pF_n(f,x+\Delta u_j(f_p^*-f))\right)+F_n^{cross},\end{aligned} \tag{14}$$

where the matrix M^p comprises only values $\{0,1\}$ and $\sum_{p=1}^{N_p}M^p=\mathbf{1}$. Here $\mathbf{1}$ represents a matrix filled entirely with the value 1. Fig. 2 illustrates the concept of decomposing a focal stack with multiple points into multiple focal stacks, each containing a single point. Combining with the affine theorem [49] and the convolution theorem [50] of Fourier transform, the formula could be written as

$$\begin{aligned}&EFS_{kn}\\ =&\frac{1}{k}EFS_n+\frac{k-1}{k}\Delta EFS\\ =&\underbrace{\frac{1}{k}EFS_n}_{Original\ term}\\ &+\frac{1}{k}\sum_{j=1}^{k-1}\sum_{p=1}^{N_p}\underbrace{FT_{2D}(M^p)}_{Occlusion\ term}*\left(\underbrace{e^{2\pi i\boldsymbol{x}_0^{p\top}\boldsymbol{A}^{-\top}\boldsymbol{\omega}}}_{Depth\ term}\underbrace{EFS_n\left(\boldsymbol{A}^{-\top}\boldsymbol{\omega}\right)}_{Shearing\ term}\right),\\ &+\underbrace{FT_{2D}(F_n^{cross})}_{Cross\ term}\end{aligned} \tag{15}$$

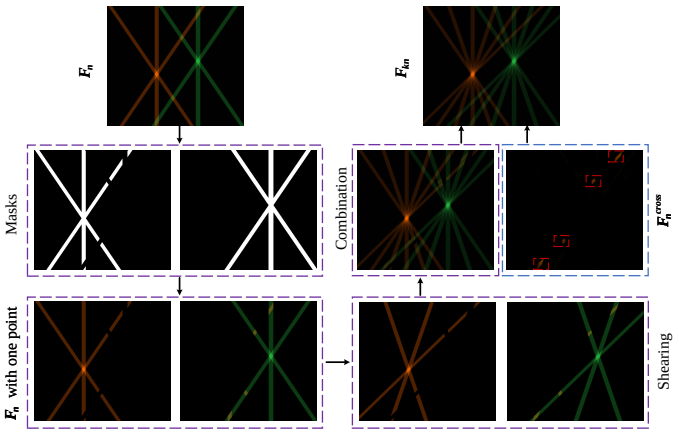


Fig. 2. Illustration of representing F_{kn} from F_n when multiple points exist in the scene. To enhance the visualization of F_n^{cross}, we adjust the contrast of the blue box (middle rightmost) to make the cross-term visible.

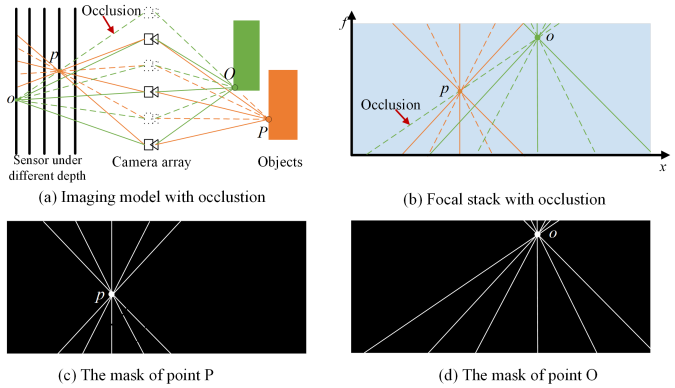


Fig. 3. Occlusion analysis for the EFS shearing. (a) is the image model with occlusion under different focal layers. We assume each constituent camera in the array operates as a pinhole camera, with each ray representing an angular sample of the scene. (b) is the focal stack of points P and O. P and O are occluded in some views. (c) and (d) are the masks of points P and O, respectively.

where $*$ represents the convolution operator.

Shared shearing operation. It is worth noting that implementing the shearing process (Eq. 14) in the spatial domain is challenging due to the coupling of the shearing operation with depth. Fortunately, in the Fourier domain, the depth and shearing operations are decoupled. Since all points share the same affine transform matrix $\boldsymbol{A}$, the shearing process can be achieved by applying the same affine transform $\boldsymbol{A}^{-\top}$ to the EFS. Compensating for depth can then be accomplished in the phase component of the EFS.

Depth as phase modulation. According to Eq. 15, the depth information $\boldsymbol{x}_0^p$ of each point p does not influence the power spectrum of the EFS and only works on the phase spectrum (*i.e.*, the depth term).

Occlusion analysis. Apart from the depth and the shearing terms, the Fourier transform of the mask M^p indicates the occlusion information of each point. Fig. 3 illustrates the generation and appearance of the occlusion in the focal stack. In Fig. 3(a), there is no occlusion in the three input views (the solid cameras and lines), however, the occlusion appears when the other three views (the dashed cameras and lines) are synthesized, *i.e.*, the orange rectangle is occluded by the green one in the top view. Fig. 3(b) shows the corresponding focal stack. Due to the occlusion of point P by O, the line passing through p and o is colored green, which matches the color of point o. Based on these observations, Fig. 3(c) and 3(d) illustrate the masks of points p and o, respectively. It is worth noting that the mask of point p comprises only 5 views, while o exhibits 6 views, indicating occlusion between p and o. Consequently, the occlusion could be obtained from the masks $\{M^p\}_{p=1}^{N_p}$.

Cross-view effects. Additionally, a cross term is introduced in Eq. 15. Fig. 2 demonstrates the cross-view phenomenon in the focal stack. Compared with the original term and the shearing term in Eq. 15, the number of cross-view points is much smaller than the total number of pixels. We have conducted several experiments to evaluate the order of magnitude of the term F_n^{cross}, and find that F_n^{cross} takes over $2\% \sim 8\%$ of the whole focal stack. For more detailed information, please refer to the supplementary material.

In summary, it is concluded that,

Proposition 1. *Given two focal stacks F_{kn} and F_n formed from densely and sparsely sampled LFs respectively, F_{kn} could be represented by F_n and the phase modulated sheared F_n in the Fourier domain, i.e., the original and sheared EFS_n.*

4 Shear-EFS-Based Dense Reconstruction

Based on the transformation model between the sparse EFS_n with only n views and the dense EFS_{kn} with kn views (Eq. 15), a two-step learning-based framework is proposed to achieve the dense LF reconstruction. In the first step, the shearing operation is applied to the input EFS_n for obtaining the shearing term in Eq. 15. In the second step, the original EFS_n and its sheared version are accumulated to obtain the $Coarse{-}EFS_{kn}$. Then, the $Coarse{-}EFS_{kn}$ is fed into a neural network to compensate for the spectrum. The framework of the proposed dense LF reconstruction is shown in Fig. 4. Previous works [26], [27] take the original EFS_n as the input and the network focuses on learning all four terms simultaneously, *i.e.*, the cross, occlusion, depth, and shearing terms in Eq. 15. Different from [26], [27], the proposed ODU-Net does not require learning the shearing term in Eq. 15 through the network, which alleviates the difficulty of training and allows for faster convergence. As a result, the proposed method achieves better spectrum compensation effectively and efficiently.

4.1 Shearing on under-sampled EFS

Given an under-sampled EPI, an aliased focal stack is obtained by applying the shearing operation following Eq. 1(a) (see Fig. 4(a)), then a 2D Fourier transform operation is applied to get the under-sampled EFS_n using Eq. 1(b) (see Fig. 4(b)). To reconstruct EFS_{kn} with k times more views than the input EFS_n, the shearing operation is applied $k-1$ times according to the shearing term in Eq. 15. Then, the original input EFS_n is combined with its sheared version to produce the $Coarse{-}EFS_{kn}$ (see Fig. 4(c)).

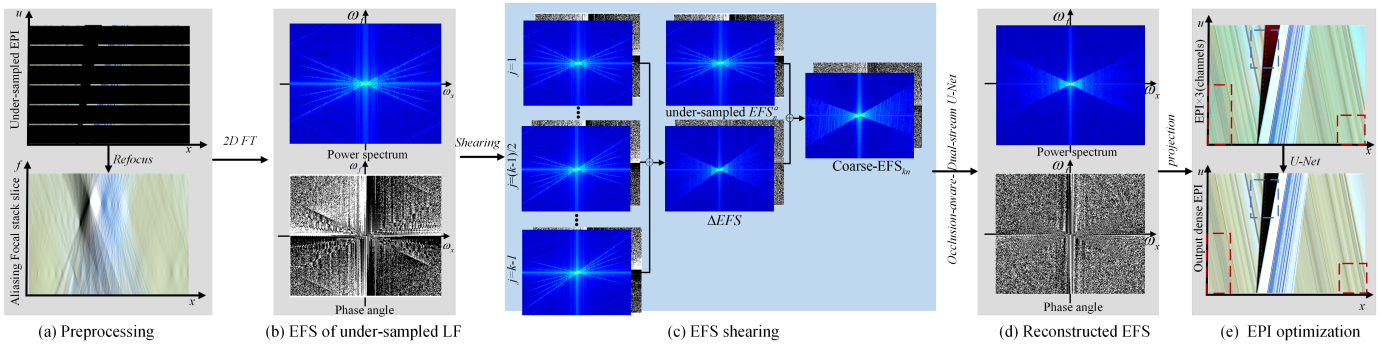


Fig. 4. The pipeline of the proposed shear-EFS-based dense LF reconstruction. The whole procedure consists of preprocessing, shear operation, EFS reconstruction, projection, and final optimization of the reconstructed EPI.

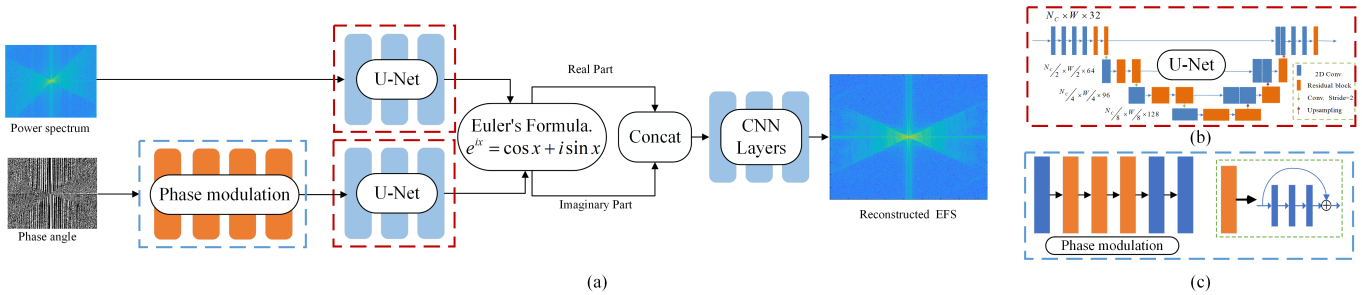


Fig. 5. (a) The architecture of the Occlusion-aware-Dual-stream U-Net (ODU-Net) for dense EFS reconstruction. (b) represents the U-Net architecture adopted in [27]. (c) provides a detailed view of the phase modulation module.

4.2 EFS reconstruction

In this section, a network named Occlusion-aware-Dual-stream U-Net (ODU-Net) is designed (see Fig. 5) to optimize $Coarse-EFS_{kn}$. According to Eq. 15, the depth acts as a phase modulation $e^{2\pi i \boldsymbol{x}_0^{p\top} \boldsymbol{A}^{-\top} \boldsymbol{\omega}}$ to the sheared EFS_n. Based on this observation, a phase modulation module (depicted as the blue dashed box in Fig. 5(a) and Fig. 5(c) for details) is first introduced to optimize the phase spectrum. Additional shortcut paths (illustrated as blue arrows in Fig. 5(c)) are incorporated to relay information forward through residual connections, which have proven useful in accelerating the training speed of the network [51], [52].

Then, considering the uncertain occlusion and cross-view effects discussed in Sec. 3.3.2, we employ a CNN to adaptively learn the filter operator for deriving the mask M^p (see Fig. 3(c) and Fig. 3(d)) and the cross term shown in Eq. 15. Then the EFS reconstruction can be formulated as the following unconstrained problem,

$$\hat{EFS}_{kn} = \Phi_{\theta_1^*}\left(\Psi_{\theta_2^*}\left(\textstyle\sum_{j=0}^{k-1} EFS_n(\boldsymbol{A}^{-\top}\boldsymbol{\omega})\right)\right), \quad (16)$$

$$(\theta_1^*, \theta_2^*) = \arg\min_{\theta_1,\theta_2} \left|EFS_{kn} - \hat{EFS}_{kn}\right| + \lambda loss_s, \quad (17)$$

where $\Psi(\cdot)$ refers to the phase modulation module in the proposed network, and $\Phi(\cdot)$ models the occlusion term and the cross term in Eq. 15. The parameters θ_1 and θ_2 correspond to the optimization targets with Φ and Ψ, respectively. The scalar λ is set to 1.5 for balancing the contributions of the two loss terms.

The first term $|EFS_{kn} - \hat{EFS}_{kn}|$ quantifies the MAE (mean absolute error) between the reconstructed EFS and the ground truth dense EFS. The second term $loss_s$ enforces the preservation of the conjugate symmetry within the reconstructed EFS,

$$loss_s = \frac{1}{N_f W} \sum_{i=0}^{N_f-1} \sum_{j=0}^{W-1} |EFS(\omega_i, \omega_j) - EFS(-\omega_i, -\omega_j)|. \quad (18)$$

Please refer to [26], [27] for the conjugate symmetry of EFS.

The architecture of the utilized neural network is illustrated in Fig. 5. Similar to the method adopted in [26], this network employs a dual-stream U-Net to extract features separately from the power spectrum and the phase spectrum, respectively. Nonetheless, before extracting features from the phase spectrum, a phase modulation module is introduced to optimize the phase spectrum obtained from the shear operation. Throughout the feature extraction process, filters are adaptively trained to acquire diverse scene point masks M^p, thus disentangling scene occlusions. Subsequently, the real and imaginary components are combined using Euler's formula to generate the real and imaginary parts. Finally, the real and imaginary parts are concatenated and forwarded to a convolutional neural network layer for optimization (Please refer to the supplementary material of [27] for the details of U-Net and CNN layers).

4.3 EPI refinement

Given the optimized $\hat{EFS}_{kn}$ from the ODU-Net, the corresponding EPI spectrum $\mathcal{E}_{kn}(\omega_u, \omega_x)$ could be directly obtained by reversing the operations in Eqs. 2(a) and 2(b). By applying a 2D IFT (inverse Fourier transform) to $\mathcal{E}_{kn}(\omega_u, \omega_x)$, one can obtain the densely-sampled EPI $\hat{E}_{kn}(u, x)$.

However, due to the interpolation operation for constructing $\mathcal{E}_{kn}(\omega_u, \omega_x)$ from $\hat{\mathcal{F}}_{kn}(f, \omega_x)$, the 'tailing' effects

(shown in the red boxes in Fig. 4(e)) and color distortion (shown in the blue boxes in Fig. 4(e)) appear, especially in the marginal views which are far away from the reference view. Hence, an additional U-Net with a perceptual loss is used to refine $\hat{E}_{kn}(u, x)$ (please refer to the supplementary material of [27] for details).

The complete shear-EFS-based dense LF reconstruction algorithm is given in Algorithm 1. H represents the height of the sub-aperture image.

Algorithm 1

Input:

An under-sampled LF with n views and disparity range d_{range}.

Output:

The reconstructed dense LF with $k * n$ views.

1: **for** $i = 1$ to H **do**
2: Obtain the EPI E_n.
3: Get the under-sampled EFS_n using Eq. 1.
4: Get the $Coarse{-}EFS_{kn}$ via the shear operation.
5: Reconstruct $\hat{EFS}_{kn}$ using the proposed ODU-Net (Fig. 5).
6: Obtain $\hat{E}_{kn}$ by inversely operating Eqs. 2(a) and 2(b).
7: Refine $\hat{E}_{kn}$ using an additional U-Net.
8: **end for**
9: Output the reconstructed dense LF with $k * n$ views.

5 Evaluations

We conduct experiments on both synthetic and real-world LF datasets [53] to evaluate our proposed shear-EFS-based dense LF reconstruction method. The real-world LF datasets are captured by both the camera array and the plenoptic camera (Lytro Illum [54]). We mainly compare our approach with five state-of-the-art learning-based methods, Wu 2019 [55], Wu 2021 [56] (without explicit depth estimation), LLFF [57] (MPI-based), Guo 2023 [16] and the EFS-without-shear [27] (EFS w/o shear). It can be observed that Wu 2021 [56], LLFF [57], Guo 2023 [16] and EFS-without-shear [27] are retrained on our training date using the released training code for a fair comparison. For Wu 2019 [55] without the original code being released, we use the trained model provided by the authors.

Quantitative evaluations are performed by measuring the average PSNR and SSIM metrics over the synthetic views of the luminance channel. In the ablation experiments, we analyze the impact of the number of shearing operations, the bound of downsampling, and the effect of the phase modulation module, respectively.

5.1 Datasets and implementation details

During the training process, both synthetic LFs (rendered by POV [58]) and real-world LFs are utilized. The synthetic dataset comprises 12 LFs containing complex textured structures, rendered using the automatic LF generator [36], [58]. Among these, 7 LFs are allocated for training, while 5 are reserved for testing. The real-world dataset, obtained from the high-resolution LF dataset by Guo *et al.* [38], consists of 26 LFs. Among these, 20 are allocated for training, while the remaining 6 are designated for testing. To illustrate the relationship between views and EFS lines and explore the impact of the number of shearing operations, we conduct experiments using the first 200 views. Additionally, we evaluate the proposed method's performance on previously unseen scenes captured by virtual camera array (generated by the Blender software [59]) and a real camera array (Disney [60]) to assess its generalization capabilities. Tab. 2 lists the parameters of all datasets. In the dense LFs, the disparity between two adjacent views for most scenes is less than one pixel, while in several scenarios, the disparity reaches two pixels. Our training data includes EPI samples obtained under different downsampling rates.

TABLE 2
Parameters of the LF datasets.

Dataset	Angular Res.	Spatial Res.
POV-Syn. LFs [58]	1×200	512×512
Blender-Syn. LFs [59]	1×128	512×512
Real LFs [38]	1×200	376×512
Couch [60]	1×101	628×1024
Church [60]	1×101	670×1024
Bike [60]	1×51	670×1024
Statue [60]	1×151	670×1024

TABLE 3
The average PSNR and SSIM values of Fig. 6

	10× Downsampling				15× Downsampling			
Shearing times	5	10	15	20	5	10	15	20
PSNR↑	34.02	38.65	40.56	40.83	31.43	31.96	38.55	39.79
SSIM↑	0.895	0.940	0.962	0.965	0.814	0.919	0.948	0.949

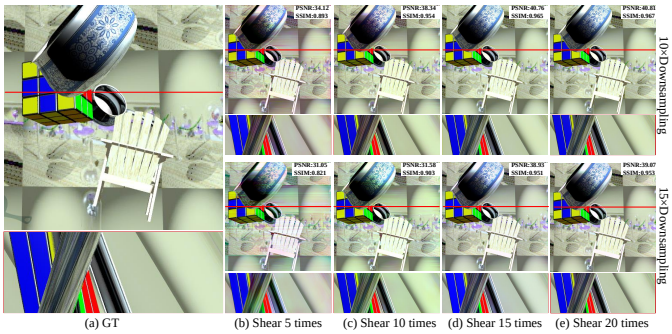
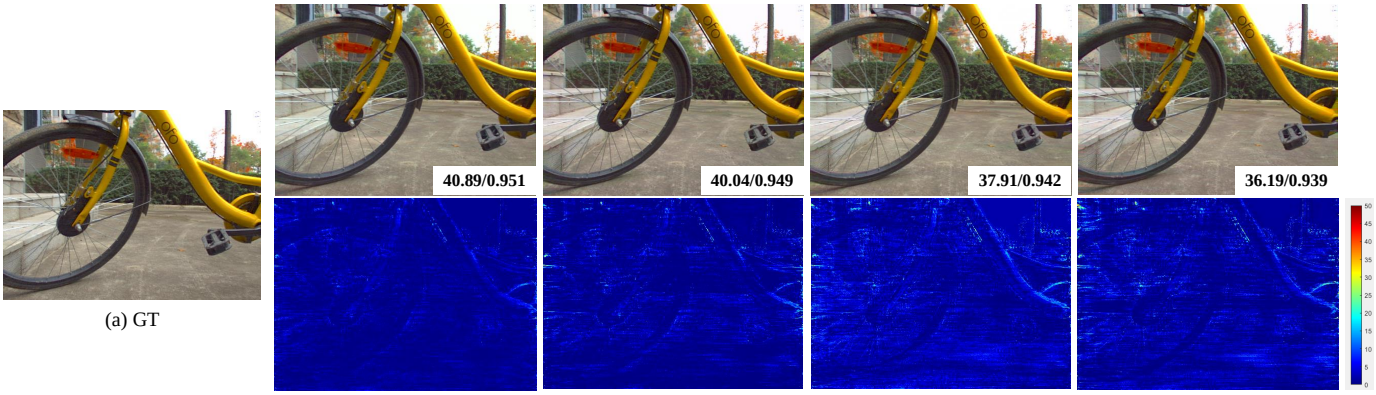


Fig. 6. Reconstruction results under different numbers of shear operations and downsampling rates. The top row shows the reconstructed views and EPIs under 10× downsampling, and the bottom row shows the results under 15× downsampling. (a) GT. (b)-(e) show the results corresponding to 5, 10, 15, and 20 shearing operations, respectively.

5.2 Ablation experiments

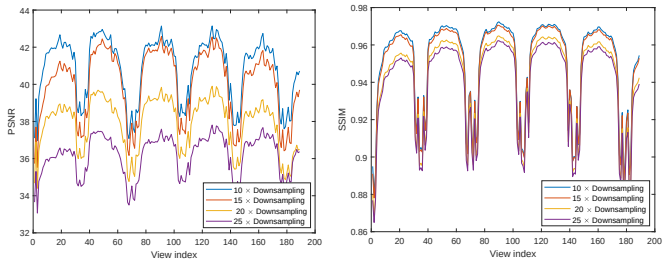
This section empirically validates the influences of the number of shear operations, the sampling rate on the reconstructed EPI, and the effect of the phase modulation module by performing the following ablation experiments.

Shearing operations analysis. In this experiment, we use the synthetic 'Pot' LF scene consisting of 200 views (POV-Synthetic LFs [58]) for testing. We conduct shearing



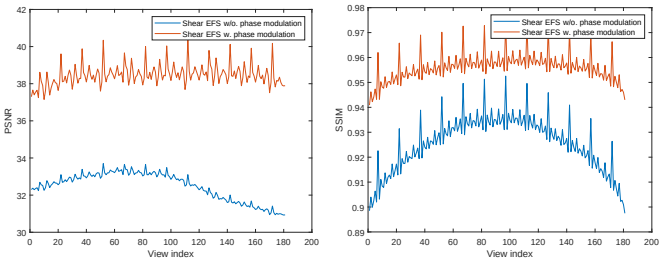
(a) GT (b) 10× Downsampling (c) 15× Downsampling (d) 20× Downsampling (e) 25× Downsampling

Fig. 7. Reconstructed views and error maps on the 'bicycle' scene under different downsampling rates.



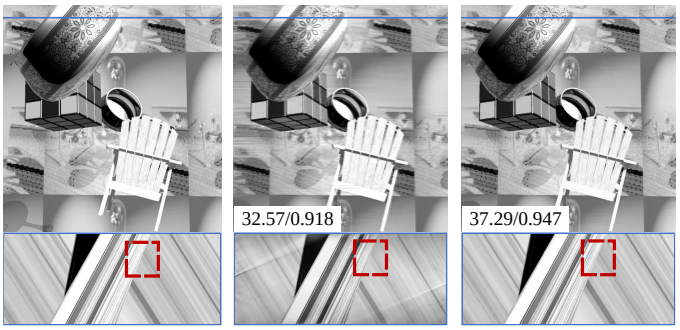
(a) PSNR (b) SSIM

Fig. 8. Quantitative comparisons (PSNR/SSIM) of each reconstructed view on the 'bicycle' scene under different downsampling rates. When implementing $k\times$ downsampling, the required number of shearing operations is $k-1$.



(a) PSNR (b) SSIM

Fig. 10. Quantitative evaluations (PSNR/SSIM) of the 'pot-cube' scene with and without the phase modulation module at each view (15× downsampling and 14 shearing operations).



(a) GT (b) Shear_EFS w/o phase modulation (c) Shear_EFS with phase modulation

Fig. 9. Reconstructed views and EPIs of the 'pot-cube' scene with or without the phase modulation module in the luminance channel. (a) represents the GT image and EPI. (b) depicts the result without phase modulation. (c) illustrates the result with phase modulation (15× downsampling and 14 shearing operations).

operations different times (5, 10, 15, and 20), considering downsampling levels of $10\times$ and $15\times$ respectively. Notably, at a $15\times$ downsampling rate, the maximum disparity between two adjacent views can reach up to 15 pixels.

As shown in Fig. 6(b), insufficient shearing operations result in reconstructed views exhibiting significant color aberration due to the large loss in the EFS. Comparing Fig. 6(d) and 6(e), it's important to note that the improvement in reconstruction performance may not always coincide with an increase in the number of shearing operations. When the number of shearing operations equals the downsampling rate (resulting in a disparity between adjacent views less than ± 1), further increasing the number of shearing operations has little effect on the reconstruction quality but escalates the algorithm's time complexity. Table 3 presents the quantitative results of Fig. 6. As a result, in the subsequent experiments, we perform $k-1$ shearing operations when dealing with $k\times$ downsampling.

Bound analysis. To evaluate the robustness of our method, we conduct experiments under $10\times$, $15\times$, $20\times$, and $25\times$ downsampling settings on a 'bicycle' real LF [38], respectively. As shown in Fig. 7(b) and 7(c), for the $10\times$ and $15\times$ downsampling settings, the proposed method could reconstruct the views with clear boundaries. Despite the reduction in PSNR/SSIM values with $20\times$ downsampling, the reconstructed results continue to display noticeable occluded edges visually. At $25\times$ downsampling, there is a significant decrease in the quality of the reconstructed view and the error map. Fig. 8(a) and 8(b) show quantitative comparisons (PSNR/SSIM) of all revamped views under different downsampling rates. Please note that the quasi-periodic 'valleys' shown in Fig. 8 result from imperfect alignment during the capture of LFs [38].

The phase modulation analysis. Here we verify the effectiveness of the phase modulation depicted in Fig.5. Fig. 9 presents a quantitative comparison of view reconstruction results on the 'pot-cube' scene, showcasing the impact of a phase modulation module versus without it (with $15\times$ downsampling and 14 shearing operations). Comparing Fig. 9(b) and 9(c), we observe a significant increase in

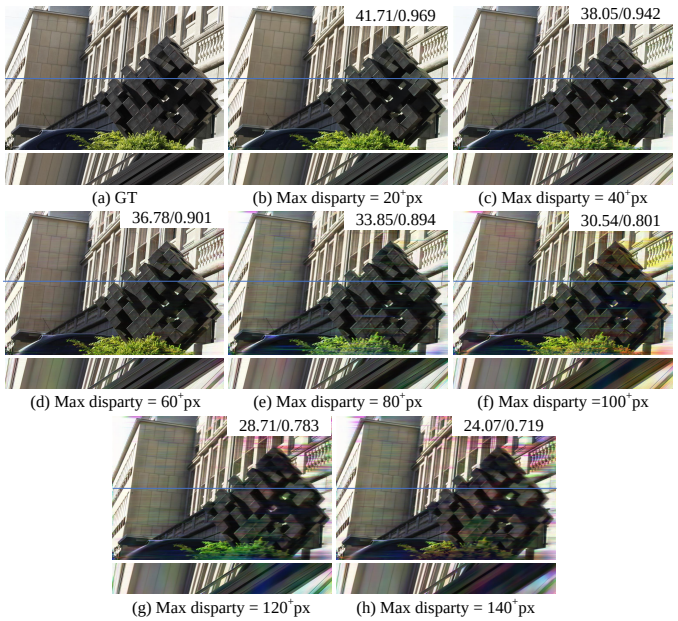


Fig. 11. Reconstruction results on the 'statue' scene [60] under different disparity ranges (pixels, px). (a) GT. (b)-(f) represent the reconstructed results with maximum disparities ranging from 20^+px to 140^+px, respectively. As the ground truth depth is unavailable for real data captured by the camera array, the labeled max disparity presented here is an estimated value. The symbol $^+$ is appended to denote this estimation.

TABLE 4
Average PSNR and SSIM values on the 'statue' scene [60] under varying disparity ranges.

Max disp. (px)	20^+	40^+	60^+	80^+	100^+	120^+	140^+
PSNR	40.91	38.27	36.20	34.23	31.09	28.46	24.32
SSIM	0.964	0.952	0.921	0.904	0.859	0.775	0.712

reconstruction quality after adding the phase modulation module. Specifically, as shown in the EPI, maintaining both the luminance of reconstruction and the consistency across views is challenging without the phase modulation module, evident in the artifact in the red box of Fig. 9(b). Fig. 10 illustrates the quantitative evaluation curves depicting reconstruction results on the 'Pot-cube' scene with and without phase modulation at each view. The results indicate a noticeable reduction in both PSNR and SSIM values for the reconstructed views when the phase modulation module is absent.

Maximum disparity. To assess the upper limit of disparity handling by our method, we downsampled the 'statue' scene of the Disney dataset [60] by factors from $10\times$ to $70\times$. At this point, the maximum disparity ranged from over 20 pixels to over 140 pixels (the input views are also changed from 15 views to 3 views). Fig. 11 provides reconstruction results on the 'statue' scene [60] under different disparity ranges. As depicted in Fig. 11, the quality of view reconstruction decreases as the disparity increases. At the disparity exceeding 100 pixels, image quality noticeably deteriorates, characterized by increased artifacts. However, as depicted by the EPI structures in Fig. 11 (b)-(f), the proposed method exhibits the ability to maintain view consistency to a considerable extent, even when confronted with large disparities. Tab. 4 displays the average PSNR and SSIM measurements on the 'statue' scene [60] across various disparity ranges, demonstrating the robustness of our method in different disparity scenarios.

TABLE 5
Average PSNR and SSIM values with different losses.

	POV-Syn. LFs [53]	Blend-Syn. LFs [53]	Real LFs [38]
MAE	34.59/0.905	35.22/0.862	35.47/0.923
MAE+SSIM	35.81/0.934	37.38/0.874	35.98/0.941
MAE+SSIM+Percep.	38.27/0.972	40.67/0.918	38.21/0.972

The loss function in EPI refinement. Our loss function of the EPI refinement U-Net consists of three parts: MAE loss, SSIM loss, and Perceptual loss. To assess the impact of these three terms on the experimental results, we conduct separate training with different combinations of them. The models are then tested on the POV-Syn. LFs [53], Blend-Syn. LFs [53], and Real LFs [38] datasets, with $15\times$ downsampling and 14 shearing operations. The combinations of loss functions and test results are presented in Tab. 5. From the last row of Tab. 5, it is evident that the perceptual loss significantly enhances the reconstruction performance.

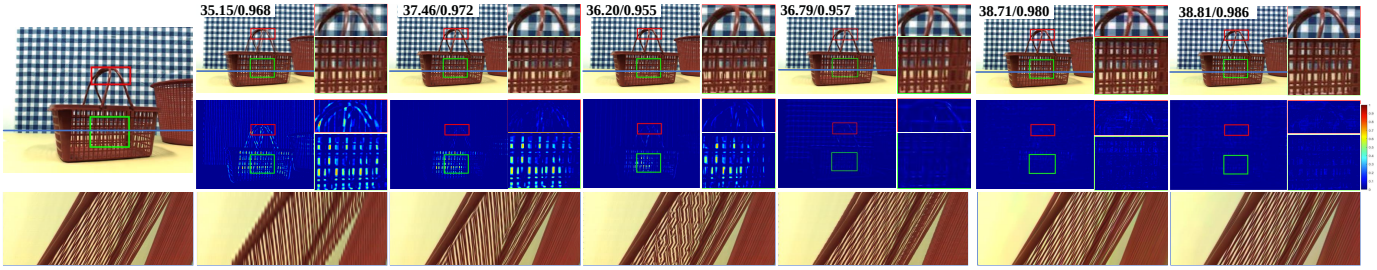
5.3 Comparisons with SOTAs

We compare our method against Wu 2019 [55], Wu 2021 [56], LLFF [57], Guo 2023 [16] and EFS-without-shear [27]. Table 6 shows the average PSNR/SSIM/LPIPS [61] measurements on both synthetic and real LFs. Qualitative comparisons among different methods on several test scenes are shown in Fig. 12, Fig. 14 and Fig. 16 respectively.

5.3.1 Real LFs captured with a plenoptic camera

We evaluate the proposed approach using real-world LF datasets [38] under $15\times$ downsampling, which contains massive static scenes in the real world. Fig. 12 shows the qualitative results on the 'basket' scene under $15\times$ downsampling, and the scene contains several thin structures, such as the basket handle. Note that, with a $15\times$ downsampling rate, the number of shearing operations is 14.

In Fig. 12(b), ghosting artifacts are apparent around the basket handle in the reconstruction result by Wu 2019 [55], which are caused by the limited receptive field of their network. Also, the Gaussian convolution kernel is only effective for small disparities.Due to the reconstruction network in Wu 2021 [56] generating multiple "plausible" results using different shear amounts, ambiguity arises when these outputs are fed into the subsequent fusion net. This ambiguity results in the failure to reconstruct regions with thin and repetitive patterns, as observed in the green box of Fig. 12(c). The MPI-based LLFF [57] tends to assign high opacity to incorrect layers in areas with ambiguous or repetitive textures, or in regions with moving content between input images. This behavior causes floating or blurred patches around repeating slender structures, as evident in the green box of Fig. 12(d). As the method of Guo 2023 [16] requires additional optical flow estimation, errors in

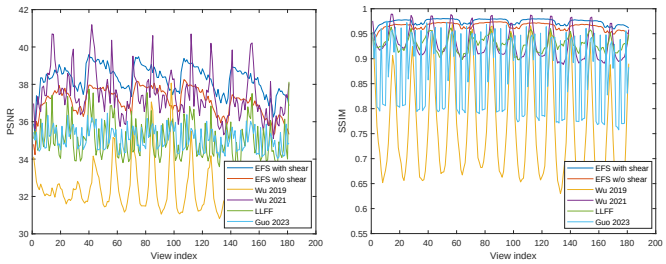


(a) GT (b) Wu 2019 (c) Wu 2021 (d) LLFF (e) Guo 2023 (f) EFS w/o shearing (g) EFS with shearing

Fig. 12. Comparisons of reconstruction results on the 'basket' scene (15× downsampling). The results consist of one reconstructed view, the error map, and the EPI of the reconstructed LF by different methods. Zooming in for better visualization. From left to right: (a) GT, the results by (b) Wu 2019 [55], (c) Wu 2021 [56], (d) LLFF [57], (e) Guo 2023 [16], (f) EFS w/o shearing [27], and (g) our method (EFS with shearing).

TABLE 6
Quantitative comparisons with SOTAs under different downsampling rates on both synthetic and real-world LFs.

		POV-Syn. LFs [53]	Blend-Syn. LFs [53]	Real LFs [38]	Couch [60]	Church [60]	Bike [60]	Statue [60]
	Downsampling	15×	15×	15×	10×	10×	5×	10×
	Shearing times	14	14	14	9	9	4	9
Wu 2019 [55]	PSNR↑	34.77	36.09	34.92	32.01	32.64	31.13	32.63
	SSIM↑	0.813	0.799	0.825	0.724	0.721	0.708	0.698
	LPIPS↓	0.093	0.064	0.078	0.113	0.08	0.110	0.098
Wu 2021 [56]	PSNR↑	37.83	38.13	37.48	34.69	32.78	32.29	34.96
	SSIM↑	0.957	0.911	0.931	0.745	0.895	0.81	0.895
	LPIPS↓	0.059	**0.036**	0.065	0.109	0.045	0.079	0.052
LLFF [57]	PSNR↑	36.46	39.42	37.02	37.25	38.85	35.51	38.53
	SSIM↑	0.922	0.898	0.925	0.916	0.962	0.875	0.956
	LPIPS↓	0.089	0.038	0.079	0.084	0.051	0.082	0.049
Guo 2023 [16]	PSNR↑	36.37	37.22	35.41	34.02	36.78	35.63	34.05
	SSIM↑	0.904	0.887	0.866	0.851	0.803	0.825	0.762
	LPIPS↓	0.084	0.045	0.075	0.083	0.079	0.091	0.073
EFS w/o shearing [27]	PSNR↑	37.45	37.67	37.74	43.05	37.95	36.77	40.82
	SSIM↑	0.952	0.903	0.938	0.928	0.964	0.939	0.959
	LPIPS↓	0.088	0.046	0.072	0.079	0.06	0.085	0.041
Ours	PSNR↑	**38.27**	**40.67**	**38.21**	**44.51**	**43.45**	**38.53**	**40.91**
	SSIM↑	**0.972**	**0.918**	**0.972**	**0.935**	**0.977**	**0.951**	**0.964**
	LPIPS↓	**0.047**	0.042	**0.058**	**0.061**	**0.031**	**0.063**	**0.032**



(a) PSNR (b) SSIM

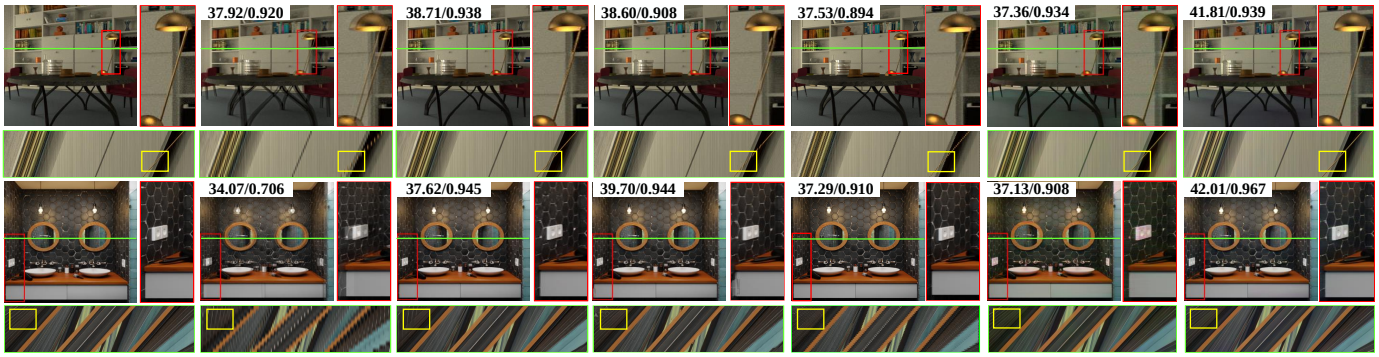
Fig. 13. PSNR and SSIM measurements for each reconstructed view on the 'basket' scene (15× downsampling and 14 shearing operations).

this estimation process may affect view consistency. This impact is noticeable in areas such as the green box and EPI of Fig. 12(e). Fig. 12(f) shows the result of EFS w/o shearing [27], where the absence of shearing operations before spectrum reconstruction results in spectral energy loss, leading to color aberration in the reconstructed views. In comparison, the proposed shear-EFS-based reconstruction method yields clearer boundaries under repetitive textures and reduces color distortion (as shown in Fig. 12(g)). Fig. 13(a) and 13(b) show the PSNR and SSIM measurements for each reconstructed view on the 'basket' scene under 15× downsampling. Overall, our method outperforms the SOTAs. The 5th column of Table 6 showcases quantitative comparisons, including PSNR, SSIM, and LPIPS, further affirming the superiority of the proposed method.

5.3.2 *Synthetic LF datasets generated by Blender*

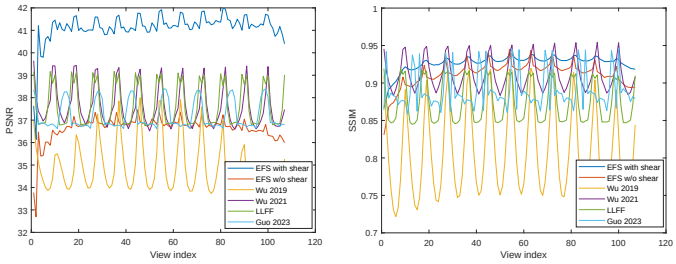
We also evaluate the proposed approach using our Blender synthetic LF datasets under 15× downsampling (the number of shearing operations is 14), which are with larger disparities (the maximum disparity is up to 15px).

The qualitative results on two synthetic LFs under 15× downsampling are shown in Fig. 14. We can see that severe ghosting artifacts occur in the results by Wu 2019 [55], and the reconstructed views are inconsistent (see the EPI

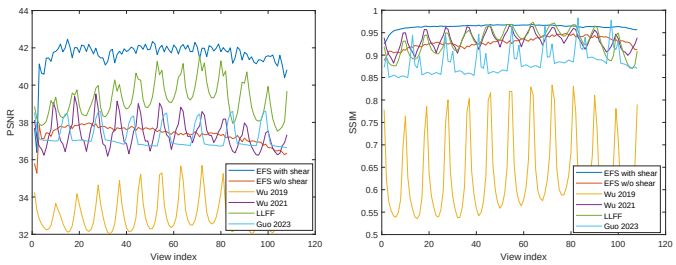


(a) GT (b) Wu 2019 (c) Wu 2021 (d) LLFF (e) Guo 2023 (f) EFS w/o shearing (g) EFS with shearing

Fig. 14. Qualitative comparisons of reconstructed views on the 'living-room' and 'washbasin' scenes (15× downsampling). The results consisi of one reconstructed view and corresponding EPI by different methods. Zooming in for better visualization. From left to right: (a) GT, the results by (b) Wu 2019 [55], (c) Wu 2021 [56], (d) LLFF [57], (e) Guo 2023 [16], (f) EFS w/o shearing [27] and (g) our method (EFS with shearing).



(a) PSNR/SSIM of the 'living-room' scene



(b) PSNR/SSIM of the 'washbasin' scene

Fig. 15. PSNR and SSIM measurements for each reconstructed view on the (a) 'living-room' and (b) 'washbasin' scenes (15× downsampling and 14 shearing operations).

of Fig. 14(b)). Similarly, a fuzzy phenomenon appears in the results by LLFF [57] on the desk lamp and the edge of the sink (the red box of Fig. 14(d)). Wu 2021 [56] cannot maintain the disparity consistency well in the case of slender structures and larger disparities (see the yellow box on EPI in Fig. 14(c)). Optical flow estimation errors result in incorrect warping, leading to erroneous reconstruction at the edges of objects in Guo 2023 method [16] (refer to the enlarged red boxes in Figure 14(e)). Even through the EFS-without-shearing method [27] reconstructs the dense LF in the frequency domain, it still suffers from color aberration caused by spectral energy errors (see Fig. 14(f)). Our method not only exhibits reduced sensitivity to spatial contents but also integrates an occlusion awareness module within the network, enabling the production of high-quality and view-consistent reconstructions (see Fig. 14(g)). As shown in Fig. 15, the PSNR/SSIM measurements for each reconstructed view achieved by our method are higher compared to the SOTA methods.

5.3.3 *Real-world LFs captured with a camera array*

To verify the effectiveness of our method under a wide baseline, we further evaluate the proposed approach using the Disney LFs [60], which are captured by a camera array.

Fig. 16 shows the results on Church LFs [60] (15× downsampling and 14 shearing operations) with wide baselines and complex occlusions (with a maximum disparity of up to 20px). Due to the limited receptive field of the network, the reconstructed view by Wu 2019 [55] exhibits serious artifacts (see Fig. 16(b)). In Wu 2021 [56], the EPI is sheared using different disparities. However, the shear operation might introduce errors, particularly under large disparities, resulting in a loss of view consistency in the reconstructed results (see the red box and EPI of Fig. 16(c)). LLFF [57] requires substantial memory to build MPI, creating a trade-off between image resolution and the layers of MPIs employed. This trade-off leads to performance degradation, especially in high-resolution input areas with large disparities, as shown in the zoomed-in rectangles and EPI in Fig. 16(d). We retrain the Guo 2023 method [16] using our datasets, however, it continues to face challenges in reconstructing views at arbitrary positions within light field datasets. Notably, in scenarios with large parallax, significant content inconsistencies emerge across views (refer to the EPI in Figure 16(e)). Additionally, errors in optical flow estimation during the preprocessing stage of the Guo 2023 method [16] have hindered the accurate reconstruction of fine, elongated texture structures within the scene, leading to noticeable errors in depicting the electric wire, as evident in Fig. 16(e). Due to the loss of spectral energy caused by large disparities, the reconstructed view of EFS w/o shearing [27] also exhibits color aberration. In contrast, our proposed method demonstrates superior performance under large disparities and exhibits better view consistency compared to other methods. This is attributed to the EFS's shearing operation and occlusion analysis incorporated within our method. Fig. 17 shows the quantitative comparison curves (PSNR/SSIM) of each reconstructed viewpoint on the 'church' scene (10× downsampling), in which we find the quantitative comparison curve of our method is significantly higher than the SOTA methods. Quantitative results on four Disney LFs are listed in the rightmost four columns of Table 6.

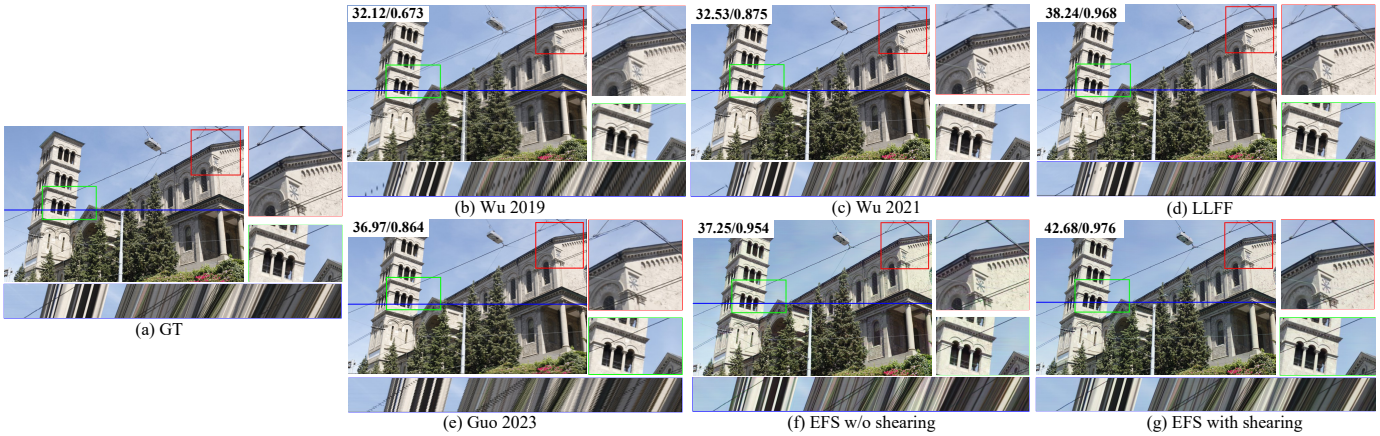
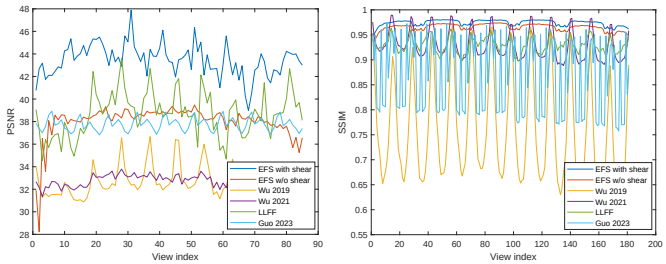


Fig. 16. Comparisons of reconstruction results on the camera array LF dataset (10× downsampling). The results consist of one reconstructed view and corresponding EPI by different methods. Several local areas are zoomed in for better visualization. (a) GT, the results by (b) Wu 2019 [55], (c) Wu 2021 [56], (d) LLFF [57], (e) Guo 2023 [16], (f) EFS w/o shearing [27] and (g) our method (EFS with shearing).



(a) PSNR (b) SSIM

Fig. 17. PSNR and SSIM measurements for each reconstructed view on the camera array LF dataset (10× downsampling and 9 shearing operations).

5.3.4 Minimum number of input views in 4D light fields

To analyze the proposed method's requirements regarding the number of input views and its feasibility for 4D LF reconstruction, Fig. 18 demonstrates the results for reconstructing a 4D LF with 9×9 views from an initial input of 2×2 views. We follow the strategy of 'horizontal first, vertical second' [27] for reconstructing the 4D LF.

The construction of EFS through spatial refocusing implies that the proposed shear-EFS-based method remains viable even with a limited number of input views, as long as refocusing can be executed (using at least two views). In Fig. 18, a significant disparity variation is evident in the foreground area, particularly in the origami crane. Compared to other SOTAs, our approach can still achieve superior reconstruction results in regions with such significant disparity variations (as observed in the zoom-in box of Fig. 18). Fig. 18(f) and (g) depict the PSNR and SSIM values for reconstructed views using different methods. It's evident that our method provides higher PSNR/SSIM values compared to the other methods.

5.4 Limitations

The theoretical analysis in this paper is based on the Lambertian assumption, which posits consistent textures for the same spatial point under different views, resulting in straight EPI lines within the scene. However, when non-Lambertian materials are present in the scene and their surfaces are rough, the EPI lines are no longer straight. Consequently, the focal stack constructed using Eq. 1 and its corresponding EFS will not exhibit the central symmetric structure depicted in Fig. 1. Therefore, our method encounters challenges when dealing with non-Lambertian scenes exhibiting such characteristics.

Fig. 19 illustrates the focal stack construction and EPI reconstruction results using our method in two distinct non-Lambertian scenes. In Fig. 19(a), we observe a non-Lambertian scene with curved objects. Specifically, the surface of the ceramic plate is uneven, leading to inconsistent textures reflected on the plate surface from different views (left). This inconsistency disrupts the linearity of the EPI line at the green line position (top right). The focal stack, constructed from this 15× downsampled EPI (middle right), and its corresponding EFS no longer exhibit the characteristics as described in [26], [27] (refer to the yellow box in Fig. 19(a)). Consequently, the shearing operation, as detailed in Sec. 3.3, becomes inapplicable. In such a scenario, our method results in errors in both the structural and color aspects of the reconstructed EPI (bottom right).

Fig. 19(b) portrays a non-Lambertian scene featuring smooth surfaces. Specifically, within the center of this scene, two mirrors are present, both with smooth surfaces. Consequently, the textures reflected on the mirror surfaces from different views remain consistent (left), resulting in the EPI line at the green line position forming a continuous straight line (top right). The focal stack constructed from this 15× downsampled EPI (middle right) and its corresponding EFS exhibit the characteristics as described in [26], [27] and are in alignment with our method. Consequently, the reconstructed EPI maintains its structural characteristics (bottom right).

Although the proposed method outperforms the SOTA methods in both view reconstruction quality and cross-view consistency preservation, the shearing operation may incur extra time costs. This problem can be mitigated by embedding shearing operations into the network [56]. Furthermore, the proposed method, relying on 2D EPI, lacks an explicit constraint for consistency from row to row in sub-aperture images, as depicted in the zoomed-in boxes in Fig. 20. A potential solution could involve utilizing 3D EFS

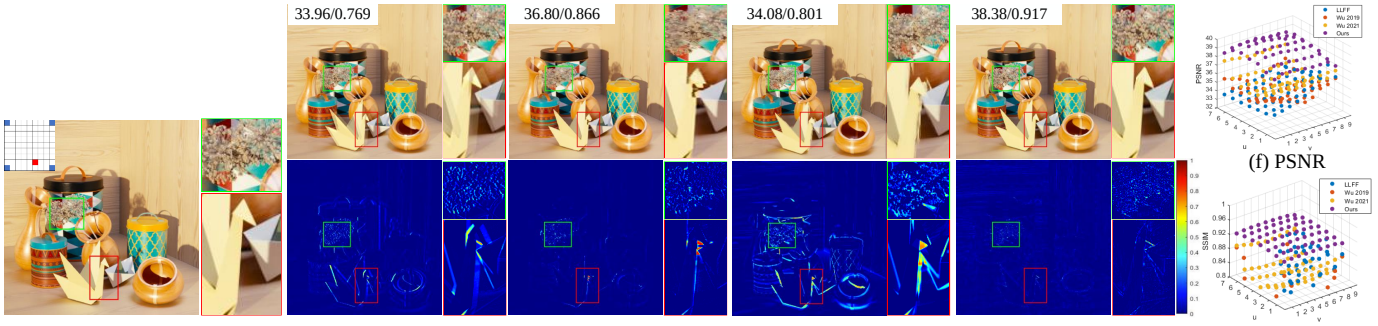


Fig. 18. Comparisons of reconstruction results on the 4D Origami LF [62] (from 2×2 to 9×9 views). The results include one reconstructed view and its corresponding error map compared against the ground truth, generated by different methods. Several local areas are zoomed in for better visualization. (a) GT, the results by (b) Wu 2019 [55], (c) Wu 2021 [56], (d) LLFF [57] and (e) our method. (f) and (g) provide the PSNR and SSIM values for 9×9 reconstructed views using different methods, respectively.

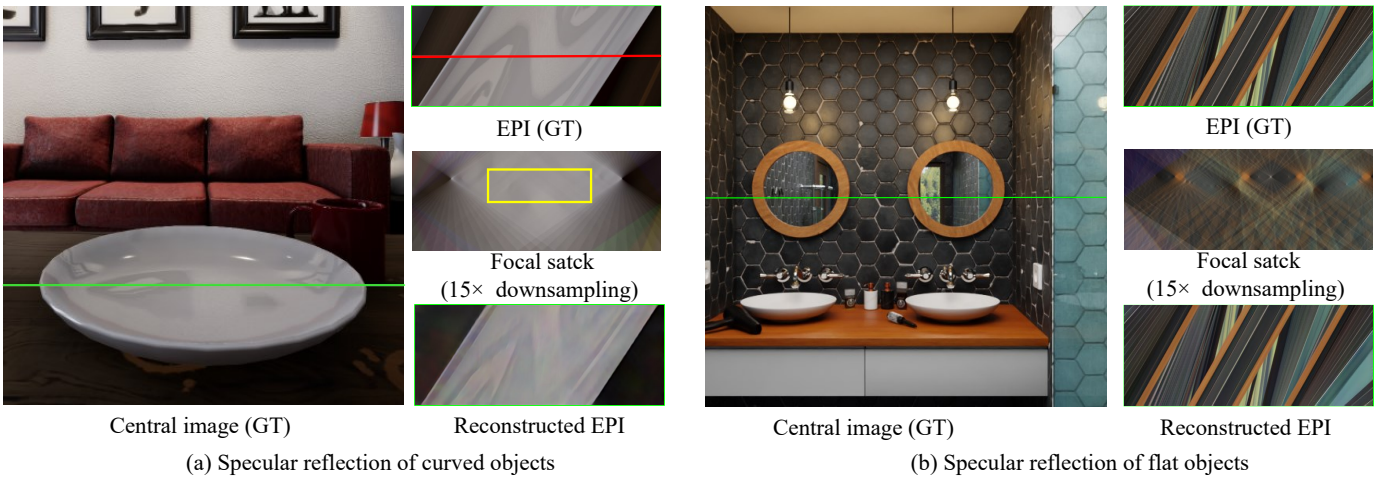


Fig. 19. Focal stack construction and EPI reconstruction results using our method in two distinct non-Lambertian scenes, depicted in (a) for scenes with curved objects[1], and in (b) for scenes featuring smooth objects.

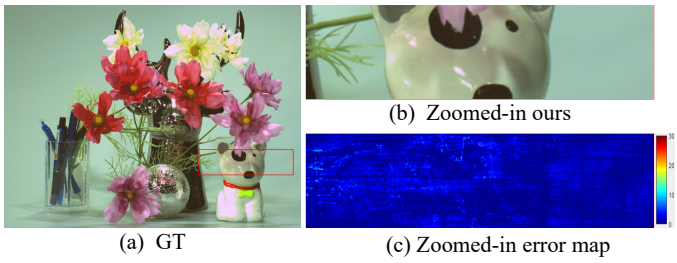


Fig. 20. An example of inconsistency across rows. (a) is the GT image. (b) is a zoomed-in version generated by our method. (c) is the zoomed-in error map. Notably, stripe-like differences can be observed across rows when comparing these images to the GT image. The scene is sourced from the dataset [63].

within the f-x-y space.

6 Conclusions

Based on the Fourier affine and convolution theorems, we analyze the relationship between the EFSs of sparse and dense LFs and make a formal breakdown of the EFS on a sparsely sampled EFS. We also analyze the occlusion in the focal stack and the EFS shearing operation and provide the occlusion model. Based on the theoretical analysis, we design a specially phase-modulated ODU-Net for reconstructing a dense LF from an undersampled LF. The proposed method exhibits superior performance under challenging conditions, such as significant disparities and complex occlusions, and maintains cross-view consistency. Experimental results have verified that the shearing strategy not only improves the accuracy of EFS completion but also reduces the complexity of network learning.

1. The central red line seen in the GT EPI is derived from the authors' released images [57]. We omit this view during the downsampling process, resulting in the absence of the red line in the reconstructed EPI.

7 Acknowledgements

The authors would like to thank anonymous reviewers for their valuable feedback.

References

[1] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow, "Deep blending for free-viewpoint image-based rendering," *ACM TOG*, vol. 37, no. 6, pp. 257:1–257:15, 2018.
[2] C. Richardt, P. Hedman, R. S. Overbeck, B. Cabral, R. Konrad, and S. Sullivan, "Capture4VR: From VR photography to VR video," in *SIGGRAPH Courses*, 2019. [Online]. Available: https://richardt.name/Capture4VR/

[3] K. Sugita, K. Takahashi, T. Naemura, and H. Harashima, "Focus measurement on programmable graphics hardware for all in-focus rendering from light fields," in *IEEE Virtual Reality*. IEEE, 2004, pp. 255–256.

[4] H. Zhu, Q. Zhang, and Q. Wang, "4D light field superpixel and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6384–6392.

[5] N. Khan, Q. Zhang, L. Kasser, H. Stone, M. H. Kim, and J. Tompkin, "View-consistent 4D light field superpixel segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7811–7819.

[6] R. Raskar and J. Tumblin, *Computational photography: mastering new techniques for lenses, lighting, and sensors*. AK Peters, Ltd., 2009.

[7] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. K. Nayar, and C. Intwala, "Spatio-angular resolution tradeoffs in integral photography," *Rendering Techniques*, vol. 2006, no. 263-272, p. 21, 2006.

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

[9] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *ACM SIGGRAPH*, 1996, pp. 43–54.

[10] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *ACM TOG*, vol. 32, no. 3, pp. 30:1–30:12, 2013.

[11] E. Penner and L. Zhang, "Soft 3D reconstruction for view synthesis," *ACM TOG*, vol. 36, no. 6, pp. 235:1–235:11, 2017.

[12] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, 2014.

[13] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4D RGBD light field from a single image," in *IEEE ICCV*, 2017, pp. 2262–2270.

[14] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM TOG*, vol. 35, no. 6, pp. 193:1–193:10, 2016.

[15] M. Guo, J. Jin, H. Liu, and 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.

[16] M. Guo, J. Hou, J. Jin, H. Liu, H. Zeng, and J. Lu, "Content-aware warping for view synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[17] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *ACM TOG*, vol. 37, no. 4, pp. 65:1–65:12, 2018.

[18] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," vol. 38, no. 4, pp. 29:1–29:14, 2019.

[19] J. Shi, X. Jiang, and C. Guillemot, "Learning fused pixel and feature-based view reconstructions for light fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2555–2564.

[20] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang, "3D photography using context-aware layered depth inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8028–8038.

[21] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *ACM SIGGRAPH*. ACM, 2000, pp. 307–318.

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

[23] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM TOG*, vol. 34, no. 1, pp. 12:1–12:13, 2014.

[24] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 133–147, 2018.

[25] M. Le Pendu, C. Guillemot, and A. Smolic, "A Fourier disparity layer representation for light fields," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5740–5753, Nov 2019.

[26] Y. Li, X. Wang, H. Zhu, G. Zhou, and Q. Wang, "Deep anti-aliasing of whole focal stack using slice spectrum," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1328–1340, 2021.

[27] Y. Li, X. Wang, H. Zhu, G. Zhou, and Q. Wang, "Dense light field reconstruction based on epipolar focus spectrum," *Pattern Recognition*, vol. 140, p. 109551, 2023.

[28] R. Ng, "Fourier slice photography," in *ACM TOG*, vol. 24, no. 3. ACM, 2005, pp. 735–744.

[29] M. Levoy and P. Hanrahan, "Light field rendering," in *ACM SIGGRAPH*. ACM, 1996, pp. 31–42.

[30] M. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *IEEE ICCV*, 2013, pp. 673–680.

[31] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *IEEE ICCV*, 2015, pp. 3487–3495.

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

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

[34] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *IEEE ICCV Workshops*, 2015, pp. 24–32.

[35] R. Ng, "Digital light field photography," Ph.D. dissertation, Stanford University, 2006.

[36] H. Zhu, M. Guo, H. Li, Q. Wang, and A. Robles-Kelly, "Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 6, pp. 3019–3033, 2021.

[37] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared epi structure for light field reconstruction," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3261–3273, 2019.

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

[39] T. Sakamoto, K. Kodama, and T. Hamamoto, "A study on efficient compression of multi-focus images for dense light-field reconstruction," in *2012 Visual Communications and Image Processing*. IEEE, 2012, pp. 1–6.

[40] M. Rizkallah, X. Su, T. Maugey, and C. Guillemot, "Geometry-aware graph transforms for light field compact representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 602–616, 2019.

[41] X. Su, M. Rizkallah, T. Maugey, and C. Guillemot, "Graph-based light fields representation and coding using geometry information," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4023–4027.

[42] X. Lv, X. Wang, Q. Wang, and J. Yu, "4D light field segmentation from light field super-pixel hypergraph representation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 9, pp. 3597–3610, 2021.

[43] P. P. Srinivasan, R. Tucker, J. T. Barron, R. Ramamoorthi, R. Ng, and N. Snavely, "Pushing the boundaries of view extrapolation with multiplane images," in *IEEE CVPR*, 2019, pp. 175–184.

[44] T. Maugey, A. Ortega, and P. Frossard, "Graph-based representation for multiview image geometry," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1573–1586, 2015.

[45] M. Rizkallah, X. Su, T. Maugey, and C. Guillemot, "Graph-based transforms for predictive light field compression based on super-pixels," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1718–1722.

[46] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker, "Deepview: View synthesis with learned gradient descent," in *IEEE CVPR*. IEEE, 2019, pp. 2367–2376.

[47] R. Tucker and N. Snavely, "Single-view view synthesis with multiplane images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 551–560.

[48] M. Levoy, *Volume rendering using the fourier projection-slice theorem*. Computer Systems Laboratory, Stanford University, 1992.

[49] R. Bracewell, K.-Y. Chang, A. Jha, and Y.-H. Wang, "Affine theorem for two-dimensional fourier transform," *Electronics Letters*, vol. 29, no. 3, p. 304, 1993.

[50] R. C. Gonzales and R. E. Woods, "Digital image processing," 2002.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

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

[53] CVPG@NWPU, "Vision and computational photography group," http://www.npu-cvpg.org/opensource, 2020.

[54] Lytro, "Lytro redefines photography with light field cameras," http://www.lytro.com, 2011.
[55] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on epi and extended applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1681–1694, 2019.
[56] G. Wu, Y. Liu, L. Fang, and T. Chai, "Revisiting light field rendering with deep anti-aliasing neural network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5430–5444, 2022.
[57] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM TOG*, vol. 38, no. 4, pp. 29:1–29:14, 2019.
[58] POV-ray, http://www.povray.org.
[59] Blender, https://www.blender.org.
[60] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM TOG*, vol. 32, no. 4, pp. 73:1–73:12, 2013.
[61] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
[62] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*. Springer, 2017, pp. 19–34.
[63] S. Moreschini, F. Gama, R. Bregovic, and A. Gotchev, "Civit dataset: Horizontal-parallax-only densely-sampled light-fields," in *Proc. Eur. Light Field Imag. Workshop*, vol. 6, 2019, pp. 1–4.