Light Field Reconstruction With a Hybrid Sparse Regularization-Pseudo 4DCNN Framework

LI FANG\(^1\), WEI ZHONG\(^1\), (Member, IEEE), LONG YE\(^2\), (Member, IEEE), RAN LI\(^1\), AND QIN ZHANG\(^2\)

\(^1\)Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China, Beijing 100024, China
\(^2\)State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China

Corresponding author: Wei Zhong (wzhong@cuc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61971383, Grant 61631016, and Grant 61801440; and in part by the Fundamental Research Funds for the Central Universities under Grant 2018XNG1825 and Grant YLSZ180226.

ABSTRACT

The densely-sampled light field (LF) is highly desirable in many applications, such as 3D reconstruction, post-capture refocusing and virtual/augmented reality. However, it is costly and challenging to capture them because of the high dimensional nature. Existing view synthesis methods employ depth information for the densely-sampled LF reconstruction from the undersampled LF with a large disparity range, but fail to achieve non-Lambertian performance. In this article, a novel coarse-to-fine LF reconstruction method is proposed. We will develop a hybrid reconstruction framework that fuses model-based sparse regularization with deep learning. Specifically, sparse regularization with directional filter bank is utilized to solve the large baseline problem and gives a coarse densely-sampled LF. In addition, for those that cannot be recovered by classical model-based methods due to limited angular resolution, is estimated by learning a pseudo 4D convolutional neural network, and thereby offering a refinement on the intermediate LF. Extensive experiments demonstrate the superiority of our method on both real-world and synthetic LF images when compared with state-of-the-art methods.

INDEX TERMS

Light field, convolutional neural network, sparse regularization, view synthesis, image-based rendering.

I. INTRODUCTION

As a promising representation of 3D visual scenes, the light field (LF) records light rays through every point traveling in every direction in the free space. The LF image contains almost all appearance information of the scene, which facilitates a wide range of applications, such as image refocusing [1], depth estimation [2], [3], 3D reconstruction [4], virtual/augmented reality [5], to just name a few. In these applications, densely-sampled LF, with high resolution in the angular domain, is highly required for sufficient information to avoid ghosting effects.

However, in many practical cases to sample a real-world scene with sufficient number of cameras to directly obtain a densely-sampled LF is challenging. For example, typical camera array systems [6] are bulky and expensive and thus are unsuitable for most commercial uses; the computer-controlled gantry [7] is always time-consuming and limited to static scenes. The commercial hand-held LF cameras such as Lytro [8] and Raytrix [9] make it convenient to acquire LF images. Unfortunately, due to restricted sensor resolution, they must make a trade-off between spatial and angular resolution.

Therefore, the required number of views has to be generated from the given sparse set of images by using view synthesis technology. Modern view synthesis methods can be divided into two categories based on their difference in prior assumptions. The first category estimates the scene depth, and then synthesizes novel views using various methods, such as soft views rendering [10], or learning-based prediction [11]. However, these depth-dependent view synthesis methods inevitably rely on the accuracy of depth information, which tends to produce artifacts where inaccurate depth estimation usually happens, such as occluded regions, non-Lambertian surfaces, etc. The second category uses specific priors such as sparsity in transformation domain for dense reconstruction [12]–[15]. They try to resolve the wavefront set of the underlying plenoptic function with specific transforms, such
as shearlets [16]. However, due to the limited angular resolution, the corresponding inverse problem is severely ill-posed and renders it very difficult to robustly recover specific parts of the target, such as occlusion and non-Lambertian regions. With recent development of deep learning solutions for visual modeling, some learning-based methods [18]–[21] have been proposed. However, most of these either need to be retrained for different interpolation rates, or suffer from aliasing when the input LF is extremely undersampled, i.e. the samples have large baselines. Moreover, at this time, neural networks are mostly used as black boxes that are for instance trained for direct inversions, or for a replacement of iteration steps in optimization algorithms. For the inverse problem of limited angle computed tomography, Bubba et al. [17] proposed a hybrid reconstruction framework that fuses model-based sparse regularization with data-driven deep learning. They utilize sparse regularization with shearlets for splitting the data into a visible part, recovered by classical model-based methods, and an invisible part, inferred in the shearlet domain by a neural network.

In this article, inspired by [17], a novel coarse-to-fine method is proposed to reconstruct densely-sampled LFs from undersampled LFs. The proposed method first utilizes sparse regularization to robustly recover the part of the wavefront set that are recoverable by classical model-based methods, referred to as recoverable part, and yield an intermediate densely-sampled LF. During the process instead of using shearlets we use our previously presented directional filter bank [22] so that a decent result can be achieved without precise disparity estimation. Therefore, our method is robust to depth inaccuracy. The remaining part, which cannot be recovered by classical model-based methods with finite iterations and is referred to as unrecoverable part, is then predicted with an encoder-like view refinement network. This will ensure a maximal amount of reliability and interpretability of our results, since deep leaning is only used to infer the unrecoverable part. A brief recoverability analysis will be given in Section III-B. Additionally, in order to make full use of 4D LF data while circumventing high computational burden, 3D convolutional neural network (CNN) is applied on stacked epipolar plane images (EPIs) in the row or column pattern. Moreover, since the sparse regularization module can handle different interpolation rates and the network only learns the unrecoverable part from the output of the former, our system can perform different interpolation rates with one model. Experimental results demonstrate the superior performance of our method over the other state-of-the-art densely-sampled LF reconstruction methods on both real-world and synthetic light field datasets with various disparity ranges.

The rest of this article is organized as follows. Section II comprehensively reviews existing methods for view synthesis and densely-sampled LF reconstruction. Section III presents the proposed approach. In Section IV, extensive experiments are carried out to evaluate the performance of the proposed approach. Finally, Section V concludes this article.

II. RELATED WORK

Many studies have focused on densely-sampled LF reconstruction using undersampled LF with a small set of views. These methods can be modeled as a reconstruction of the plenoptic function using limited samples. We divide the related work into two categories: those that use depth estimation, and those that do not.

A. DEPTH-DEPENDENT VIEW SYNTHESIS

These approaches typically first estimate the scene depth and then warp the given images to the novel view based on the estimated depth, where the depth information works as correspondence map for view reprojection. A number of depth estimation methods have been developed specifically for stereo images [23], and for multiview images as well [3], [4], [24]–[28].

In recent years, many learning-based methods have been proposed for depth image-based view synthesis. Flynn et al. [11] proposed a deep learning method to synthesize novel views using a stereo pair or sequence of images with wide baselines. Srinivasan et al. [29] proposed to synthesize a 4D LF image from a 2D RGB image based on estimated 4D ray depth. Kalantari et al. [19] proposed to synthesize novel sub-aperture images with two sequential networks that perform depth estimation and color prediction successively. Zhou et al. [30] and Mildenhall et al. [31] trained a network that infers alpha and multiplane images. The novel views are synthesized using homography and alpha composition. Jin et al. [32] proposed a coarse-to-fine framework that accepts irregular input with large baseline. However, all these approaches are based on the Lambertian assumption without explicitly addressing the non-Lambertian challenge.

B. DEPTH-INDEPENDENT LF RECONSTRUCTION

Densely-sample LF reconstruction can be considered as consecutive reconstruction (interpolation) of the underlying plenoptic function from its given incomplete measurements (input views). The required bounds for sampling the LF of a scene have been defined in [33]. For a light field that the disparity between neighboring views is less than one pixel, novel views can be produced using linear interpolation [33].

For undersampled LF, direct interpolation will cause ghosting effects in the rendered views. To mitigate this effect, some studies have investigated LF reconstruction with specific priors. Levin and Durand [34] exploit dimensionality gap priors to synthesize novel views from a set of images sampled in a circular pattern. Shi et al. [13] sampled only the boundary viewpoints or diagonal viewpoints to recover the full light field using sparsity analysis in the Fourier domain. However, Fourier and other isotropic transforms have been found less efficient for representing images. The sought transforms should be local, directional and multiresolution [35]. Among the candidates, shearlets have been shown to be effective in reconstructing a densely-sampled EPI from an undersampled EPI with a large disparity range. Vagharshakyan et al. [14], [15] proposed an approach using
the sparse representation of EPIs in the shearlet transform domain. Gao et al. further introduced optic flow [36], [37] and video frame interpolation methods [38] to improve the reconstruction result of shearlets-based methods. However, shearlet transform requires a precise estimation of the disparity range of the undersampled LF in order to design a shearlet system with decent scales and to pre-shear the undersampled EPIs. In addition, due to the limited angular resolution, it is difficult for classical model-based methods to recover some specific parts, such as occlusion and non-Lambertian regions.

Recently, some learning-based approaches were also proposed for depth-free reconstruction. Yeung et al. [39] applied a coarse-to-fine model using an encoder-like view refinement network for larger receptive field. Wu et al. [20] proposed a “blur-restoration-deblur” framework as learning-based angular detail restoration on 2D EPIs. Wang et al. [40] introduced an end-to-end learning-based framework using 3D light field volumes, achieving over 10× faster than Wu et al.’s framework [20]. Based on the observation that an EPI shows clear structure when sheared with the disparity value, Wu et al. [21] proposed to fuse a set of sheared EPIs for LF reconstruction. However, since each EPI is a 2D slice of the 4D LF, the accessible spatial and angular information of these EPI-based models are severely restricted.

III. THE PROPOSED APPROACH

A. LF REPRESENTATION AND PROBLEM STATEMENT

The 7D continuous plenoptic function [41] describes the set of light rays traveling in every direction through every point in 3D space, from a geometric optics perspective. In practice, the plenoptic function is simplified to its 4D version, which records light rays by their intersections with two parallel planes, namely the two-plane parameterization. Consider a pinhole camera, with focal plane \((u, v)\) and focal length \(f\), moving along the second plane \((s, t)\), an oriented light ray defined in this system first intersects the image plane at coordinate \((u, v)\), corresponding to the spatial axes, and then intersects the second plane at coordinate \((s, t)\), corresponding to the angular axes, and is thus denoted by \(L(u, v, s, t)\).

In the two-plane system, we can consider the \((s, t)\) plane as a set of cameras with their focal plane on the \((u, v)\) plane, and each camera collects the light rays leaving the \((u, v)\) plane and arriving at a point on the \((s, t)\) plane. Thus the 4D light field can be represented as a 2D array of images, such as the one shown in Fig. 1. Each image recorded by the camera is called a sub-aperture image. Let \(L(u, v, s, t) \in \mathbb{R}^{W \times H \times N \times M}\) denote a densely-sampled LF containing \(N \times M\) sub-aperture images of spatial dimension \(W \times H\), which are sampled on the angular plane with a regular 2D grid of size \(N \times M\).

By gathering the light field samples with a fixed spatial coordinate \(v\) and an angular coordinate \(t\) (or \(u\) and \(s\)), one can produce a slice \(E_{\alpha,\beta}(u, s)\) (or \(E_{\alpha,\beta}(v, t)\)), resulting a map called EPI. The EPI contains relative motion between the camera and object points. Points with different depths can be visualized as lines with different slopes in the EPI. Compared with regular photo images, an EPI has a very well-defined structure. Any visible scene point appears in one of the EPIs as a line whose slope reflects the depth of the point and the measured intensity over the line reflects the intensity of emanated light from that scene point. For Lambertian reflectance model (any point in the scene emanates light in different direction with same intensity), a point on the surface of an observed object forms a line with a constant intensity in the EPI. In Fourier domain, the frequency support of a layer with depth \(Z\) is confined to a line with a slope of \(Z\). Therefore, the spectrum of a Lambertian scene between \(Z_{\text{min}}\) and \(Z_{\text{max}}\) is constrained within a wedge bounded by two lines passing through the origin with slopes \(Z_{\text{min}}\) and \(Z_{\text{max}}\).

Since the EPI is a 2D slice of the 4D LF with more definitive structure, the reconstruction of densely-sampled LF from undersampled LF can be considered as an inpainting problem on the EPI. Without loss of generality, we assume that the undersampled EPI is a uniformly downsampled version of the densely-sampled one

\[
y = Ax + \eta, \tag{1}
\]

where \(A\) is a measuring matrix, and \(\eta\) models measurement noise. The measurements \(y\) form an incomplete EPI where only rows from the available views are presented, while everywhere else values are 0.

In order to reconstruct the densely-sampled EPI from the undersampled one (see Fig. 2), one need to solve the inverse problem described in (1). Due to the limited angular resolution, not all features of the measured object \(x\) are captured under \(A\) and the resulting inverse problem is severely ill-posed.

B. OVERVIEW OF THE PROPOSED METHOD

There exists an abundance of inversion strategies for natural images, among which shearlets-based methods [14], [15]...
have been proven to be effective. Because of the multiresolution property, shearlet transform is effective to reconstruct densely-sampled EPIs from undersampled EPIs with large baseline.

![FIGURE 2. Illustration of the inpainting problem on the EPI and examples of the recoverable part (boxed in green lines) and unrecoverable part (boxed in red lines and yellow dash line): (a) Densely-sampled EPI; (b) undersampled EPI (2 input views).](image)

However, for scenes containing occlusions or non-Lambertian surfaces, the inverse problem is severely ill-posed. Due to limited angular resolution, some part of the coefficients cannot be robustly recovered by shearlets-based methods since not enough samples are contained in the measured data. For example, as shown in Fig. 2(a), some lines are partially occluded in the densely-sampled EPI. In a task with just $2 \times 2$ input views, only the top pixels and the bottom pixels are sampled in the undersampled EPI, and the presence of occlusion makes the endpoints of some lines have different values, see Fig. 2(b). For lines bounded by two endpoints with similar values (boxed in green lines, to only list a few), they can be robustly recovered by shearlets-based methods, referred to as recoverable part, while for lines bounded by two endpoints with different values (boxed in red lines, to only list a few), cannot, referred to as unrecoverable part. In addition, lines in non-Lambertian regions are not exactly straight (see Fig. 2(a), boxed by yellow dash lines, to only list a few), making partial of the corresponding shearlet coefficients unrecoverable. Precisely this part is sought to be recovered by an inference in the shearlet domain by means of a trained neural network. But due to the high dimensional nature of the LF, it is challenging to perform learning methods in shearlet domain. Note that the above recoverability analysis is only qualitative insight inspired by [17] without rigorous proof. It will prove to be effective in Section IV through extensive experiments.

Therefore, we choose to first reconstruct the recoverable part using sparse regularization with our previously proposed directional filter banks [22], and then learn a network to estimate the unrecoverable part in the spatial domain.

The prediction of the missing part can be solved by residual learning. Such an inference of unrecoverable information from the knowledge of its recoverable counterpart is feasible, since the EPIs under the same scene geometry have a similar texture structure. Moreover, the state-of-the-art LF reconstruction networks are more effective than shearlets-based method in the case of small disparity range. Being aware of this unique characteristic as well as the great success of deep learning, we propose a learning-based approach to explore the EPI texture structure for densely-sampled LF reconstruction. It can be considered as a refinement of the coarse result reconstructed by sparse regularization from the undersampled LF, which makes our framework work in a coarse-to-fine manner.

Fig. 3 shows the flowchart of the proposed framework, which comprises three steps. The first step of our algorithm consists in solving a sparse regularization problem. Promoting sparsity in the transform domain allows to robustly recover the recoverable part. The next step is to generate an estimation of the unrecoverable part by training an artificial deep neural network. The architecture, referred to as pseudo 4DCNN, is an encoder-like network that performs inpainting on stacked EPIs. Having an estimation of the unrecoverable part at hand, the final step consists in fusing both parts in the spatial domain. In the following, the details of the proposed approach are presented step-by-step.

C. RECOVERING THE RECOVERABLE PART VIA SPARSE REGULARIZATION

Shearlet transform is extremely effective in reconstructing a densely-sampled EPI from a undersampled EPI with a large disparity range. Since we are going to process EPIs of scenes placed between the focal plane and infinity, whose spectrum locates at the wedge in the range of $[3/4\pi, \pi]$ in the frequency plane, only the shear operation with a positive sign is considered. For a more comprehensive frequency analysis of both LF and EPI, we refer the interested reader to [14], [42].

In such elaborately-designed shearlet systems, a pre-shearing by the minimum disparity is required in order to correctly process the undersampled EPI. A precise depth range should be estimated first. In our process, we utilize our previously proposed directional filter bank [22] to replace shearlets. The following wedge-shaped filters with cosine roll-off transition band are used

$$\hat{W}(\omega m, \omega l) = \begin{cases} 
\sin \frac{\omega m - (\omega m_1 - \omega_T/2)}{2\omega_T}, & \omega m_1 - \omega_T/2 \leq \omega m < \omega m_1 + \omega_T/2, \\
\cos \frac{\omega m - (\omega m_2 - \omega_T/2)}{2\omega_T}, & \omega m_2 - \omega_T/2 \leq \omega m < \omega m_2 + \omega_T/2, \\
0, & \text{else}
\end{cases}$$

(2)
where \( \omega_{m1} \) and \( \omega_{m2} \) determine the angular range that the corresponding frame covers in the frequency plane, \( L \) is the filter size, and \( \omega_T \) is the width of the transition band. The wedge-shaped filters can be placed anywhere, even in an angular range across the diagonal, as shown in Fig. 4. In contrast, classical shearlets are restricted in two cone-like regions with fixed shear operation. Since our filter bank has a more flexible partitioning, it is possible to perform better frequency tiling than classical shearlets in the context of EPI representation.

The analysis filter at scale \( j \) and direction range \( \omega \) is given by

\[
\hat{\psi}_{j, \omega} = \hat{w}_{\omega} \cdot \hat{g}_j, \tag{3}
\]

where \( \hat{g}_j \) is the Fourier transform of the highpass filter \( g_j = h_j * g_{j+1} \uparrow 2 \) at scale \( j \). \( h_j = h_{j} * h_{j+1} \uparrow 2 \) is the corresponding lowpass filter, \( 0 \leq j < J \), and \( \uparrow 2 \) means upsampling by 2. In addition, \( h_j \) and \( g_j \) are lowpass and highpass filters of the 2D two-channel nonsampled filter bank presented in [43]. The lowpass component is extracted by

\[
\hat{\phi} = h_0. \tag{4}
\]

As the frame elements are not orthogonal, one needs also the dual frame elements. They can be constructed based on pseudoinverse. First, we set

\[
\hat{\Psi} = |\hat{\phi}|^2 + \sum_{j=0}^{J-1} \sum_{\omega} |\hat{\psi}_{j, \omega}|^2. \tag{5}
\]

Then, the synthesis filters are defined in Fourier domain, as follow

\[
\hat{\phi} = \frac{\hat{\phi}}{\hat{\psi}} \cdot \hat{\psi}_{j, \omega} = \hat{\psi}_{j, \omega}. \tag{6}
\]

Since we only consider the wedge in the range of \([\frac{3}{4}\pi, \pi]\), which corresponds to the depth range from the focal plane to infinity. The resulting direct transform \( F \) for discrete values \( x_j \) and \( j = 0, \ldots, J - 1, \omega \in \frac{1}{4}\pi, \pi, m \in \mathbb{Z}^2 \) is

\[
F(x_j) = (f_j, \omega)(m) = (x_j \cdot \hat{\psi}_{j, \omega})(m),
\]

\[
f_0(m) = (x_j \cdot \hat{\phi})(m). \tag{7}
\]

The corresponding inverse transform is then

\[
F^*(f_j, \omega, f_0) = \sum_{j=0}^{J-1} (f_j, \omega * \hat{\psi}_{j, \omega})(m) + (f_0 * \hat{\phi})(m). \tag{8}
\]

The reconstruction of \( x \) given the sampling matrix \( A \) and the measurements \( y \) can be cast as an inpainting problem, with constraint to have solution which is sparse in the transform domain, i.e.,

\[
x^* = \arg \min_x \frac{1}{2} \|Ax - y\|^2 + \lambda \cdot \|F(x)\|_1. \tag{9}
\]

It is typically solved via an iterative inpainting process with \( t \) iterations. We make use of the double relaxation method from [15]:

\[
\begin{align*}
\hat{x}_i &= F^* T_{1, \lambda}(F(x_i + \alpha(x_0 - A' \circ x_i))), \\
\tilde{x}_i &= \hat{x}_i + \beta_1(\hat{x}_i - x_{i-1}), \\
x_{i+1} &= \tilde{x}_i + \beta_2(\tilde{x}_i - x_{i-2}), \tag{10}
\end{align*}
\]

where

\[
\begin{align*}
\beta_1 &= \frac{\text{sum}(x_0 - \hat{x}_i) \circ A' \circ (\hat{x}_i - x_{i-1}))}{\text{sum}(\tilde{x}_i - x_{i-1}) \circ A' \circ (\hat{x}_i - x_{i-1})}, \\
\beta_2 &= \frac{\text{sum}(x_0 - \tilde{x}_i) \circ A' \circ (\tilde{x}_i - x_{i-2}))}{\text{sum}(\tilde{x}_i - x_{i-2}) \circ A' \circ (\tilde{x}_i - x_{i-2})}. \tag{11}
\end{align*}
\]

Here, \( \text{sum}(\cdot) \) returns the sum of all the elements in the input matrix, “\( \circ \)” denotes the elementwise (Hadamard) product,
\(A'\) is a logical measuring mask obtained by rearranging \(A\), where ideally \(x_i \circ A' = x_0\), and typically \(x_0 = x\). \(T_{\lambda}(\cdot)\) is a hard thresholding operator applied on transform domain coefficients and \(\alpha\) is an acceleration parameter.

Please note that (10) does not solve the optimization problem (9) if \(F\) is not an orthonormal basis. It is a reconstruction algorithm that is only inspired by (9) and often yields a good reconstruction performance in practice. As will be demonstrated in Section IV, the proposed frame produces better results with inaccurate depth information, making our approach more robust.

D. LEARNING THE UNRECOVERABLE PART WITH CNN

Given the intermediate LF reconstructed by iterative process, we are now working on the inference of unrecoverable part by means of a CNN. The overall goal is to find a nonlinear mapping \(NN(\inter\, L) \approx \unr\, L\) from the recoverable part \(\inter\, L\) to the unrecoverable part \(\unr\, L\)

\[
NN_0(\inter\, L) \approx \unr\, L. \tag{12}
\]

Because of the high dimensional nature of the LF, it is rather difficult to work on 4D data, and individually processing all EPIs cannot make full use of the input views. Therefore, we take advantage of pseudo 4DCNN in [40] and build an encoder-like view refinement network on stacked EPIs. Taking the row volume as an example (shown in Fig. 5), both the angular and spatial correlation are employed by the 3D CNN. Since the goal is to predict the unrecoverable part which has the similar texture structure with the recoverable part, residual learning is used in this module. The overview of the pseudo 4DCNN is given in Fig. 3, where the unrecoverable part is recovered first horizontally, then vertically, and added to the recoverable part successively.

Concretely, for the refinement of the intermediate LF \(\inter(u, v, s, t)\) with size of \(W \times H \times N \times M\), we first extract 3D row volume for each column \(t^s\), \(t^s = 1, 2, \ldots, M\), and permute it as

\[
V_r(v, u, s) = \text{perm}(\inter(u, v, s, t^s)) \tag{13}
\]

with the size of \(H \times W \times N\), where perm(\(\cdot\)) denotes permutation operation. The 3D row volumes are refined by the row network modeling as \(F_{\text{row}}(\cdot)\), and then compose a new intermediate LF as

\[
L'_{\text{inter}}(u, v, s, t^s) = \text{perm}(V_r(v, u, s) + F_{\text{row}}(V_r(v, u, s))) \tag{14}
\]

Similarly, we extract 3D column volume from \(L'_{\text{inter}}(u, v, s, t)\) for each row \(s^r\), \(s^r = 1, 2, \ldots, N\), and permute it as

\[
V_c(u, v, s) = \text{perm}(L'_{\text{inter}}(u, v, s^r, t)) \tag{15}
\]

with the size of \(W \times H \times M\). The 3D column volumes are refined by the column network modeling as \(F_{\text{col}}(\cdot)\), and then compose the reconstructed LF as

\[
L^r(u, v, s^r, t) = \text{perm}(V_c(u, v, s^r) + F_{\text{col}}(V_c(u, v, s^r))) \tag{16}
\]

We use a fully convolutional encoder-decoder architecture for both proposed network \(F_{\text{row}}(\cdot)\) and \(F_{\text{col}}(\cdot)\). As depicted in Fig. 6, the network for the residual prediction consists of an encoder and a decoder. The encoder comprises three convolution layers. The first layer contains 64 filters of size \(1 \times 5 \times 5 \times 3\), where each filter operates on \(5 \times 5\) spatial regions across \(3\) adjacent EPIs. The second layer contains 16 filters of size \(64 \times 3 \times 3 \times 3\). The last layer contains 16 filter of size \(16 \times 3 \times 3 \times 3\). The decoder also comprises three convolution layers which are exactly symmetric to those of the encoder. Each layer uses a stride of 1 followed by a rectified linear unit (ReLU) except for the last one. To avoid border effects, we appropriately pad the input and feature maps before every convolution operation to maintain the input and output at the same size.

E. EDGE ENHANCEMENT LOSS FUNCTION

The proposed framework is designed to reconstruct the desired 4D LF by inpainting its EPIs. Since the EPIs have a clear texture structure, we proposed an edge enhancement loss function consisting of three parts.

The first part provides supervision for the output LF by calculating the weighted average over the entire mean squared errors (MSE) between the output LF images and ground-truth ones, i.e.,

\[
\ell_s = \frac{1}{MNHW} \sum_{s^r=1}^{N} \sum_{t^r=1}^{M} \omega_{s^r,t^r} \|L_{GT}(s^r,t^r) - L_{\text{out}}(s^r,t^r)\|_2^2, \tag{17}
\]

where \(\omega_{s^r,t^r}\) is the prior sensitive weights in [40]. Specifically, synthesized views after the row network \(F_{\text{row}}(\cdot)\) are inferred from the input views, while a portion of those views synthesized after the column network \(F_{\text{col}}(\cdot)\) only receive prior information propagated from earlier synthesized views.

In the prior sensitive scheme, more attention is paid to the errors of the later synthesized views by using larger weights. According to the order that views are generated and the level of received prior knowledge, all the synthesized views are divided into four groups and their MSE against the ground truth are summed up with corresponding weights.
The weighting coefficient $\omega_{s^*, t^*}$ for the synthesized view at $(s^*, t^*)$ is particularly set as

$$
\omega_{s^*, t^*} = \begin{cases} 
0,1, & s^* \in [1:p:N], \ t^* \in [1:p:M] \\
1, & s^* \in [1:p:N], \ t^* \notin [1:p:M] \\
1, & s^* \notin [1:p:N], \ t^* \in [1:p:M] \\
2, & s^* \notin [1:p:N], \ t^* \notin [1:p:M], 
\end{cases}
$$

(18)

where $p$ is the upscale factor.

Finally, the output reconstructed LF image is optimized by minimizing the objective

$$
\ell = \lambda_1 \ell_s + \lambda_2 \ell_{\text{row}} + \lambda_3 \ell_{\text{col}},
$$

(20)

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are the weighting for the reconstruction accuracy and the edge enhancement, which are empirically set to 1, 9 and 9, respectively.

IV. EXPERIMENTS AND RESULTS

A. DATASETS AND IMPLEMENTATION DETAILS

Both synthetic LF images from the 4D light field benchmark [44], [45] and real-world LF images captured with a Lytro Illum camera provided by Standford Lytro LF Archive [46] and Kalantari et al. [19] were employed to train and test. Specifically, 20 synthetic images and 100 real-world images were used for training, while 9 synthetic data, including 4 LF images from the HCI [44] dataset and 5 LF images from the old HCI [45] dataset, and 3 datasets with totally 70 real-world LF images captured with a Lytro Illum camera were used for test, namely 30scenes [19], Occlusions [46] and Reflective [46]. These datasets cover several important factors in evaluating the methods for light field reconstruction, including high-frequency texture, natural illumination and practical camera distortion, large baselines, occlusions and reflective surfaces. We removed the border views and cropped the LF images to $7 \times 7$ views as ground truth, and then downsampled to $2 \times 2, 3 \times 3$ views as the input.

In the procedure of sparse regularization, since no disparity information is known, the directional filter system is constructed with 4 scales, which means that the maximum disparity that can be processed is up to 16 pixels, i.e. $J = \lceil \log_2 d_{\text{max}} \rceil$. We used 3, 5, 9, 17 directions in the scales from coarser to finer, respectively. For each scale, the central frequency of these direction subbands are uniformly placed from $\frac{1}{2}\pi$ to $\pi$, resulting in a slightly wider coverage than the classical shearlets. The thresholding level $\lambda_i$ decreases with the iteration number linearly in the range [2, 0.02]. Note that these two values are for the case of using a normalized undersampled EPI $\varepsilon$, i.e.

$$
\varepsilon = \frac{\varepsilon - \min(\varepsilon)}{\max(\varepsilon) - \min(\varepsilon)},
$$

(21)

where $\max(\cdot)$ and $\min(\cdot)$ return the maximum value and the minimum value of an input matrix, respectively. The reconstructed $x$ using this normalized $\varepsilon$ is then rescaled back to original range of values via

$$
x = (\max(\varepsilon) - \min(\varepsilon)) \cdot x + \min(\varepsilon).
$$

(22)
Besides, we set the maximum iteration number $t = 20$, and the acceleration parameter $\alpha$ is 10.

Both the sparse regularization and the network are implemented with PyTorch. During training the network, small patches in the same position of each view are extracted. The spatial patch size is $48 \times 48$ and the stride is 20. Similar to other SR methods, we only processed the luminance Y channel in the YCbCr color space. Moreover, we adopted ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. And the batch size was set to 64. The filters of 3D CNNs are initialized from a zero-mean Gaussian distribution with standard deviation 0.01 and all the bias are initialized to zero. The learning rate is initially set to $10^{-4}$ and then decreased by a factor of 0.5 every 10 epochs until the validation loss converges. Other hyperparameters not mentioned were set to PyTorch default values. Note that in the process of the pseudo 4DCNN, the LF images in the range of $[0, 255]$ are automatically rescaled to $[0, 1]$ by the Pytorch PILImage methods.

B. COMPARISONS WITH STATE-OF-THE-ART METHODS

We compared with 5 state-of-the-art learning-based methods that were specifically designed for densely-sampled LF reconstruction, i.e., Wu et al. [20], Yeung et al. [39], Wang et al. [40], Wu et al. [21], Kalantari et al. [19], and Jin et al. [32]. Among these methods, the last three are geometry-based. In the subsequent content, we conducted various experiments for comparisons.

1) REAL-WORLD SCENES

As to experiments on real-world scenes, we constructed $7 \times 7$ LFs from both $3 \times 3$ and $2 \times 2$ sparse views (see Fig. 8). The performances of comparative methods [19]–[21], [39] were obtained via implementing source codes released by respective authors. While for the approaches of Wang et al. [40] and Jin et al. [32], since they have not open-sourced their software, we directly take the results in [40], [32]. Please note that since we used the exact same datasets as Jin et al. [32] for both training and test, taking their results directly for comparison is feasible. As for Wang et al. [40], because they used some private dataset, only the results on the public dataset 30scenes [19] are compared. For all these methods used for comparison, we used the best configurations listed in the corresponding references. We used the average value of PSNR and SSIM over all views in each scene to quantitatively measure the quality of reconstructed densely-sampled LFs. Due to limited space, we only report the average result for all data entries in each dataset.

We first designed the experiment $3 \times 3$ to $7 \times 7$. As to evaluation on the proposed approach, we followed the protocols in [20] and used 30scenes, as well as two representative scenes, Reflective29 and Occlusion16, from the Stanford Lytro Light Field Achieve [46]. We compared with Wu et al. [20], Yeung et al. [39] and Wang et al. [40]. Table 1 lists the corresponding results on the real-world datasets. Our proposed model performs better for all datasets than three comparing methods: with 3.53 dB, 0.16 dB, and 1.36 dB reconstruction advantage over Wu et al. [20], Yeung et al. [39] and Wang et al. [40] in terms of PSNR, respectively; and with 0.006, 0.004, and 0.002 reconstruction advantage over Wu et al. [20], Yeung et al. [39] and Wang et al. [40] in terms of SSIM, respectively.

For the task $2 \times 2$ to $7 \times 7$, we carried out comparison with the method by Kalantari et al. [19], Wu et al. [20], Wu et al. [21], Yeung et al. [39] and Jin et al. [32]. The method by Wang et al. [40] cannot be compared since their approach requires 3 views in each angular dimension to provide enough information for the transposed convolution layers. Our testing dataset contains 30scenes, Occlusions and Reflective. This test set contains 113 LFs which is sufficient to provide objective evaluation of model performance.

We used the four corner images of a LF as input. Reconstruction quality was measured with PSNR and SSIM averaged over the luma component, and over all reconstructed views. As shown in Table 2, our proposed approach obtains an average of 39.83 dB, 0.08 dB higher than that of Yeung et al. [39] in terms of PSNR, respectively. The latter achieves 39.75 dB and is the second best among all methods.

It can be observed that EPI-based methods, including Wu et al. [20] and Wu et al. [21], are inferior compared with others. The former produced satisfactory reconstruction results using $3 \times 3$ inputs but declines quickly for a sparser input ($2 \times 2$). The latter performs relatively better on 30scenes but fails on Reflective, as depth information

| Methods       | Wu et al. [20] | Yeung et al. [39] | Wang et al. [40] | Ours |
|---------------|----------------|-------------------|------------------|------|
| 30scenes      | 41.02/0.988    | 44.53/0.999       | 43.28/0.992      | 44.70/0.994 |
| Occlusion16   | 38.86/0.985    | 39.53/0.987       | -                | 39.72/0.992 |
| Reflective29  | 46.10/0.993    | 47.85/0.996       | -                | 47.91/0.997 |
| Average       | 41.11/0.988    | 44.48/0.990       | 43.28/0.992      | 44.64/0.994 |

TABLE 1. Quantitative comparisons (PSNR/SSIM) of the reconstruction quality of the proposed approach with the state-of-the-art methods under the task $3 \times 3$ to $7 \times 7$.  

FIGURE 8. Illustration of sampling patterns: (a) $3 \times 3$; (b) $2 \times 2$.  

Wu et al. [20], Yeung et al. [39], and Jin et al. [32].
TABLE 2. Quantitative comparisons (PSNR/SSIM) of the reconstruction quality of the proposed approach with the state-of-the-art methods under the task 2 × 2 to 7 × 7 on real-world scenes. The input undersampled LFs are sampled at the four corners during both training and test.

| Methods       | L. Fang et al. [19] | Wu et al. [20] | Wu et al. [21] | Yeung et al. [39] | Jin et al. [32] | Ours     |
|---------------|---------------------|----------------|----------------|-------------------|----------------|----------|
| 30 scenes     | 41.40/0.982         | 33.66/0.918    | 39.17/0.975    | 42.77/0.986       | 42.75/0.986    | 42.91/0.990 |
| Occlusions    | 37.25/0.972         | 32.72/0.924    | 34.41/0.955    | 38.88/0.980       | 38.51/0.979    | 38.93/0.986 |
| Reflective    | 38.09/0.953         | 34.76/0.930    | 36.38/0.944    | 38.33/0.960       | 38.35/0.957    | 38.37/0.962 |
| Average       | 38.59/0.969         | 33.55/0.924    | 36.23/0.957    | 39.75/0.976       | 39.59/0.975    | 39.83/0.981 |

TABLE 3. Quantitative comparisons (PSNR/SSIM) of the reconstruction quality of the proposed approach with the state-of-the-art methods under the task 2 × 2 to 7 × 7 on synthetic scenes. The input undersampled LFs are sampled at the four corners during both training and test.

| Methods | L. Fang et al. [19] | Wu et al. [20] | Wu et al. [21] | Yeung et al. [39] | Jin et al. [32] | Ours     |
|---------|---------------------|----------------|----------------|-------------------|----------------|----------|
| HCl     | 32.85/0.909         | 26.64/0.744    | 31.84/0.889    | 32.30/0.900       | 37.14/0.966    | 37.08/0.968 |
| HCl old | 38.58/0.944         | 31.43/0.850    | 37.61/0.942    | 39.69/0.941       | 41.80/0.974    | 41.75/0.971 |
| Average | 36.03/0.928         | 29.30/0.803    | 35.05/0.922    | 36.41/0.923       | 39.73/0.970    | 39.69/0.969 |

TABLE 4. Comparisons of the running time (in second) of different methods for reconstructing a densely-sampled LF.

| Methods       | L. Fang et al. [19] | Wu et al. [20] | Wu et al. [21] | Yeung et al. [39] | Jin et al. [32] | Ours     |
|---------------|---------------------|----------------|----------------|-------------------|----------------|----------|
| HCl 2 × 2 to 7 × 7 | 661.42            | 244.81         | 96.61          | 0.85              | 40.21          | 196.01   |

is utilized as guidance. The possible reason is that only 2 rows or columns of pixels are available during the reconstruction of each EPI, making it difficult to recover the intermediate linear structures, especially when the scenes are complicated. In contrast, Yeung et al. [39], Jin et al. [32] and our method employ the similarity of neighboring pixels and make a better use of the full 4D LF data, indicating that the pseudo 4D filters effectively explore the spatial and angular relations between input views. The difference is that we perform 3D CNNs on EPI volumes, while the other two use 2D CNNs on spatial and angular dimension successively.

2) SYNTHETIC SCENES

We used the 9 synthetic LF images, including 4 LF images from the HCI [44] dataset and 5 LF images from the old HCI [45] dataset. The angular resolution of these synthetic images is 9 × 9 and were cropped to 7 × 7. While for the input, we adopted the four corner images (2 × 2) of a LF as input to evaluate the performance of the proposed framework for large disparity. Table 3 shows a quantitative evaluation of the proposed approach on the synthetic dataset compared with other methods. It can be observed that Jin et al. [32] achieves the best results on the synthetic datasets. This is because the HCI dataset contains LF images with large baselines, which measure the ability on very sparse sampling, and geometry-based methods are efficient to reconstruct LFs from large baseline sampling. While our method is comparable to Jin et al. [32] on synthetic datasets, without explicitly using geometry information.

We also visually compared the reconstruction results of different algorithms, as shown in Fig. 9. It can be observed that Wu et al. [20] and Wu et al. [21] fail to recover delicate structures, such as the leaves and the textures on the wall, while Kalantari et al. [19] and Yeung et al. [39] struggle with large disparities. In contrary, our approach produces accurate estimations, which are closer to ground-truth ones. Note that since the source code of Jin et al. [32] has not been released yet, some figures are extracted from [32].

The running time (in second) of different methods for reconstructing a densely-sampled LF is lasted in Table 4. All methods were tested on a desktop with Intel i9-9900k @ 3.6 GHz CPU, 32 GB RAM and a Nvidia GeForce RTX 2080ti GPU, except for Jin et al. [32]. Note that we directly take the results in [32] since their software has not been open-sourced yet. They used a desktop with the same RAM size and GPU as ours, but the CPU is Intel i7-8700 @ 3.6 GHz, so that their method could be slightly faster when running on our platform. The reason why our method is relatively slow is that the sparse regularization is time-consuming. This is the cost of reliability and interpretability as deep learning is only used to estimate the unrecoverable part.

C. ABLATION STUDY

To better illustrate the advantage of the proposed method, we conduct ablation experiments on 30scenes with the two components of our framework, i.e., the sparse regularization module and the pseudo 4DCNN module.

1) THE EFFECTIVENESS OF THE DIRECTIONAL FILTER BANK

In the proposed method, the sparse regularization is performed by promoting sparsity on the coefficients obtained by our previously proposed directional filter bank. Alternatively, shearlet transform [14], [15] is used for such tasks. To validate the advantages of our directional filter bank, we experimented both shearlets in Vagharshakyan et al. [14], [15] and our filter bank under the task 3 × 3 to 7 × 7. Both methods used the same configuration as previously described. Since the EPIs should be pre-sheared, to avoid the influence of depth
estimation error, we looped through all possible minimum disparities using shearlets and tested our filter bank with the same best pre-shearing for each scene.

Fig. 10 depicts quantitative comparisons of the average PSNR and SSIM on 30 scenes. It is easy to find out that the proposed directional filter bank performs significantly better.
better than shearlets. Specifically, our directional filter bank achieves 37.18 dB per LF scene on average, 0.80 dB higher than shearlets used in Vagharskhyan et al. [14], [15] (36.38 dB).

2) THE EFFECTIVENESS OF THE PSEUDO 4DCNN

To demonstrate the effectiveness of the pseudo 4DCNN module, we conduct ablation experiments with variants of the pseudo 4DCNN, and Table 5 lists the results. We can see that the results decrease 0.87 dB on average without the prior sensitive weights from [40]. Also, the results decrease 1.02 dB on average without the edge enhancement loss. It can be demonstrated that each component of the pseudo 4DCNN contributes to improving the performance.

| TABLE 5. Quantitative comparisons (PSNR/SSIM) on 30 scenes with variants of the pseudo 4DCNN under the task 3 × 3 to 7 × 7. |
|---|---|
| Variants | PSNR/SSIM |
| Ours (full) | 44.70/0.994 |
| Without prior sensitive weights | 43.83/0.993 |
| Without edge enhancement loss | 43.68/0.991 |

V. CONCLUSION

In this article, a hybrid sparse regularization-pseudo 4DCNN framework is proposed to directly synthesize novel views of 4D densely-sampled LF from sparse input views. We assemble sparse regularization with directional filter bank for recovering the recoverable part and a pseudo 4DCNN for inference of the unrecoverable part to solve the problem of reconstructing LFs with large disparity range. Our framework works in a coarse-to-fine manner by first reconstructing an intermediate densely-sampled LF using our previously proposed directional filter bank, and then apply 3D CNNs on row and column EPI volumes successively for refinement. Extensive evaluations on real-world and synthetic LF scenes show that our proposed framework outperforms various state-of-the-art approaches.

REFERENCES

[1] J. Fiss, B. Curless, and R. Szeliski, “Refocusing plenoptic images using depth-adaptive splatting,” in Proc. IEEE Int. Conf. Comput. Photography (ICCP), May 2014, pp. 1–9.
[2] J. Chen, J. Hou, Y. Ni, and L.-P. Chau, “Accurate light field depth estimation with superpixel regularization over partially occluded regions,” IEEE Trans. Image Process., vol. 27, no. 10, pp. 4889–4900, Oct. 2018.
[3] S. Wanner and B. Goldluecke, “Variational light field analysis for disparity estimation and super-resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 3, pp. 606–619, Mar. 2014.
[4] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, “Scene reconstruction from high spatio-angular resolution light fields,” ACM Trans. Graph., vol. 32, no. 4, p. 73, 2013.
[5] F. C. Huang, K. Chen, and G. Wetzstein, “The light field stereoscope: Immersive computer graphics via factored near-eye light field displays with focus cues,” ACM Trans. Graph., vol. 34, no. 4, p. 60, 2015.
[6] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, “High performance imaging using large camera arrays,” ACM Trans. Graph., vol. 24, no. 3, pp. 765–776, Jul. 2005.
[7] The (New) Stanford Light Field Archive. Accessed: Aug. 1, 2020. [Online]. Available: http://lightfield.stanford.edu/acq.html

[8] Lytro Illum. Accessed: May 17, 2017. [Online]. Available: https://www.lytro.com/
[9] Raytrix. Accessed: Aug. 1, 2020. [Online]. Available: https://www.raytrix.de/
[10] E. Penner and L. Zhang, “Soft 3D reconstruction for view synthesis,” ACM Trans. Graph., vol. 36, no. 6, pp. 235:1–235:11, 2017.
[11] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, “Deep stereo: Learning to predict new views from the World’s imagery,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 5515–5524.
[12] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, “Compressive light field photography using overcomplete dictionaries and optimized projections,” ACM Trans. Graph., vol. 32, no. 4, pp. 1–11, 2013.
[13] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, “Light field reconstruction using sparsity in the continuous Fourier domain,” ACM Trans. Graph., vol. 34, no. 1, pp. 1–13, Dec. 2014.
[14] S. Vagharskhyan, R. Bregovic, and A. Gotchev, “Light field reconstruction using shearlet transform,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 1, pp. 133–147, Jan. 2018.
[15] S. Vagharskhyan, R. Bregovic, and A. Gotchev, “Accelerated shearlet-domain light field reconstruction,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 7, pp. 1082–1091, Oct. 2017.
[16] M. Genzel and G. Kutyniok, “Asymptotic analysis of inpainting via universal shearlet systems,” SIAM J. Imag. Sci., vol. 7, no. 4, pp. 2301–2339, Jun. 2014.
[17] T. A. Babba, G. Kutyniok, M. Lassas, M. Marz, W. Samek, S. Siltanen, and V. Srinivasan, “Learning the invisible: A hybrid deep learning-shearlet framework for limited angle computed tomography,” Inverse Problems, vol. 35, no. 6, Jun. 2019, Art. no. 064002.
[18] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, “Learning a deep convolutional network for light-field image super-resolution,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Dec. 2015, pp. 24–32.
[19] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, “Learning-based view synthesis for light field cameras,” ACM Trans. Graph., vol. 35, no. 6, pp. 193:1–193:10, 2016.
[20] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, “Light field reconstruction using convolutional network on EPI and extended applications,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1681–1694, Jul. 2019.
[21] G. Wu, Y. Liu, Q. Dai, and T. Chai, “Learning sheared EPI structure for light field reconstruction,” IEEE Trans. Image Process., vol. 28, no. 7, pp. 3261–3273, Jul. 2019.
[22] L. Fang, L. Ye, Y. Tie, W. Zhong, and Q. Zhang, “Design of linear-phase nonsubsampled nonuniform directional filter bank with arbitrary directional partitioning,” J. Vis. Commun. Image Represent., vol. 51, pp. 23–28, Feb. 2018.
[23] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” Int. J. Comput. Vis., vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.
[24] I. Pearson, M. Brookes, and P. Dragotti, “Plenoptic layer-based modeling for image based rendering,” IEEE Trans. Image Process., vol. 22, no. 9, pp. 3405–3419, Sep. 2013.
[25] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 2, pp. 328–341, Feb. 2008.
[26] S. N. Sinha, D. Scharstein, and R. Szeliski, “Efficient high-resolution stereo matching using local plane sweeves,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1582–1589.
[27] G. Zhang, J. Jia, T.-T. Wong, and H. Bao, “Consistent depth maps recovery from a video sequence,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 6, pp. 974–988, Jun. 2009.
[28] S. Wanner and B. Goldluecke, “Globally consistent depth labeling of 4D light fields,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 41–48.
[29] P. P. Srinivasan, T. Wang, A. Sreeval, R. Ramamoorthi, and R. Ng, “Learning to synthesize a 4D RGBD light field from a single image,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2243–2251.
[30] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, “Stereo magnification: Learning view synthesis using multiplane images,” ACM Trans. Graph., vol. 37, no. 4, p. 65, Aug. 2018.
[31] 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 Trans. Graph., vol. 38, no. 4, p. 29, 2019.
[32] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong, and J. Yu, “Flexible, fast and accurate densely-sampled light field reconstruction network,” 2019, arXiv:1909.01341. [Online]. Available: https://arxiv.org/abs/1909.01341

[33] Z. Lin and H.-Y. Shum, “A geometric analysis of light field rendering,” Int. J. Comput. Vis., vol. 58, no. 2, pp. 121–138, Jul. 2004.

[34] A. Levin and F. Durand, “Linear view synthesis using a dimensionality gap light field prior,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 1831–1838.

[35] N. Minh Do and M. Yue Lu, “Multi-dimensional filter banks and multiscale geometric representations,” Found. Trends Signal Process., vol. 5, no. 3, pp. 157–264, 2012.

[36] Y. Gao, R. Bregovic, A. Gotchev, and R. Koch, “MAST: Mask-accelerated shearlet transform for densely-sampled light field reconstruction,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Shanghai, China, Jul. 2019, pp. 187–192.

[37] Y. Gao, R. Koch, R. Bregovic, and A. Gotchev, “Fast: Flow-assisted shearlet transform for densely-sampled light field reconstruction,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Taipei, Taiwan, Sep. 2019, pp. 3741–3745.

[38] Y. Gao, R. Koch, R. Bregovic, and A. Gotchev, “IEST: Interpolation-enhanced shearlet transform for light field reconstruction using adaptive separable convolution,” in Proc. 27th Eur. Signal Process. Conf. (EUSIPCO), A Coruna, Spain, Sep. 2019, pp. 1–5.

[39] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, “Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues,” in Proc. ECCV, 2018, pp. 138–154.

[40] 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 Proc. Eur. Conf. Comput. Vis., 2018, pp. 340–355.

[41] E. Adelson and J. Bergen, “The plenoptic function and the elements of early vision,” in Computational Models of Visual Processing. Cambridge, MA, USA: MIT Press, 1991.

[42] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, “Plenoptic sampling,” in Proc. ACM SIGGRAPH, 2000, pp. 307–318.

[43] A. L. Da Cunha, J. Zhou, and M. N. Do, “The nonsubsampled contourlet transform: Theory, design, and applications,” IEEE Trans. Image Process., vol. 15, no. 10, pp. 3089–3101, Oct. 2006.

[44] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, “A dataset and evaluation methodology for depth estimation on 4D light fields,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 19–34.

[45] S. Wanner, S. Meister, and B. Goldluecke, “Datasets and benchmarks for densely sampled 4D light fields,” in Proc. VMV, 2013, pp. 225–226.

[46] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein, Stanford Lytro Light Field Archive. [Online]. Available: http://lightfields.stanford.edu/LF2016.html

**LI FANG** received the B.E. degree in electronic information engineering and the M.S. degree in communication and information system from Wuhan University, Wuhan, China, in 2010 and 2012, respectively, and the Ph.D. degree in communication and information system from the Communication University of China, Beijing, China, in 2017.

He was a Visiting Scholar with Ryerson University, Toronto, ON, Canada. He is currently a Research Assistant with the Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China. His research interests include computer vision and virtual reality.

**WEI ZHONG** (Member, IEEE) received the B.E. degree in electronic engineering and the Ph.D. degree in circuits and systems from Xidian University, in 2004 and 2010, respectively.

She joined the Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China, Beijing, China, in 2011, where she has been an Associate Professor, since 2015. Her research interests include multirate signal processing, image processing, and video affective content analysis.

**LONG YE** (Member, IEEE) received the B.E. degree in electronic engineering from Shandong University, Jinan, China, in 2003, and the M.S. and Dr. Eng. degrees from the Communication University of China, Beijing, China, in 2006 and 2012, respectively.

He was a Visiting Scholar with Ryerson University, Toronto, ON, Canada. He is currently a Professor with the State Key Laboratory of Media Convergence and Communication, Communication University of China. His research interests include computer vision, image compression, and virtual reality.

**RAN LI** received the B.E. degree in radio and television engineering from the Communication University of China, Beijing, China, in 2018, where she is currently pursuing the M.S. degree with the Key Laboratory of Media Audio and Video, Ministry of Education.

**L. Fang et al.: LF Reconstruction With a Hybrid Sparse Regularization-Pseudo 4DCNN Framework**

**QIN ZHANG** received the Ph.D. degree in engineering from The University of British Columbia, Vancouver, BC, Canada, in 1990.

He worked as a Research and Development Scientist at the EE Department, The University of British Columbia, from 1990 to 1995. In 2004, he was the Dean of the TCL Industrial Research Institute. He is currently a Professor with the State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China. His research interests include computer vision, audio and video processing, and virtual reality.

---

**Authors:**

**LI FANG**

**WEI ZHONG**

**LONG YE**

**RAN LI**

**QIN ZHANG**