Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4570755, 10 pages
https://doi.org/10.1155/2022/4570755

Research Article

Flexible Light Field Angular Superresolution via a Deep Coarse-to-Fine Framework

Qian Wang, Li Fang, Long Ye, Wei Zhong, Fei Hu, and Qin Zhang

Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education, Beijing 100024, China

Correspondence should be addressed to Li Fang; lifang8902@cuc.edu.cn

Received 10 October 2021; Accepted 25 January 2022; Published 10 March 2022

Academic Editor: Ivan Lee

Acquisition of densely-sampled light fields (LFs) is challenging. In this paper, we develop a coarse-to-fine light field angular superresolution that reconstructs densely-sampled LFs from sparsely-sampled ones. Unlike most of other methods, which are limited by the regularity of sampling patterns, our method can flexibly deal with different scale factors with one model. Specifically, a coarse restoration on epipolar plane images (EPIs) with arbitrary angular resolution is performed and then a refinement with 3D convolutional neural networks (CNNs) on stacked EPIs. The subaperture images in LFs are synthesized first horizontally, then vertically, forming a pseudo 4DCNN. In addition, our method can handle large baseline light field without using geometry information, which means it is not constrained by Lambertian assumption. Experimental results over various light field datasets including large baseline LFs demonstrate the significant superiority of our method when compared with state-of-the-art ones.

1. Introduction

The light field (LF) encodes the distribution of light into a high-dimensional function, contains rich scene visual information [1, 2] and has a wide range of applications in various fields, such as image refocusing [3], 3D scene reconstruction [4], depth inference [5], and virtual augmented reality [6]. In order to obtain high-quality views without ghosting effects, many studies have focused on dense sampling of LF [7].

Dense sampling of LF means great acquisition difficulties. Early light field cameras, such as multicamera arrays and light field racks etc. [8], are bulky and expensive in hardware. In recent years, the introduction of commercial and industrial light field cameras such as Lytro [9] and RayTrix [10] has brought light field imaging into a new era. Unfortunately, due to the limited resolution of the sensor, a trade-off must be made between spatial resolution and angular resolution.

One possible solution to this problem is view synthesis, which synthesizes novel views from a sparse set of input views. One type of existing previous work is to estimate the scene depth as auxiliary information [11, 12], but it relies heavily on the depth estimation, which tends to fail in occluded regions, as well as in glossy or specular ones. The other is based on sampling and consecutive reconstruction of the plenoptic function [13–16]. They do not use the depth information as an auxiliary mapping but suffer from either aliasing or blurring problem when the input LF is extremely undersampled.

Some learning-based methods have recently appeared [17–19], and they can be roughly classified into two categories: nondepth based and depth based. But most of them have average reconstruction quality under large parallax conditions. Besides, retraining is required for different scale factors, which increases the difficulty of actual acquisition.

In our previous work [20], we proposed a learning-based model for reconstructing densely-sampled LFs via angular superresolution, which is achieved by using an image superresolution network on epipolar plane images (EPIs). However, EPIs have very clear structure, which is very different from natural images. The performance is degraded by large-baseline sampling. In this paper, we provide a few distinguishable improvements and enable flexible and accurate reconstruction of a densely-sampled LF from very sparse

sampling. We inherit the coarse-to-fine framework in [20], that is, the proposed model consists of a coarse EPI angular superresolution module and an efficient EPI stack refinement module. Specifically, the coarse EPI angular superresolution module magnifies each EPI individually using a specially designed EPI superresolution network, where independent coefficients are used for row and column EPIs, respectively. We further refine the coarse results using 3DCNNs on stacked EPI based on photoconsistency between subaperture images. The overall frame explores pseudo 4DCNN, which is capable of making full use of LF data. In addition, we introduced perceptual loss [21] to better fit the network training. Experimental results demonstrate the superiority of our method in reconstruction with higher numerical quality and better visual effect.

The rest of this paper is organized as follows: Latest developments of view synthesis and LF reconstruction are introduced in Section 2. Our presented approach is described in Section 3. The performance evaluation is given in Section 4. Finally, Section 5 summarizes this paper.

2. Related Works

The problem of reconstructing a complete densely-sampled LF from a set of sparsely sampled images has been extensively studied. These algorithms can be divided into depth-dependent view synthesis that depends on depth information and depth-independent LF reconstruction that does not depend on depth information.

2.1. Depth-Dependent View Synthesis. Depth-dependent view synthesis approaches typically consist of two steps to synthesize the novel view of the scene [22], i.e., first estimating depth map at the novel view or the input view, and then using it to synthesize the novel [23], Kalantari et al. [17] proposed the first deep learning system for view synthesis with two sequential networks that perform depth estimation and color prediction successively. Srinivasan et al. [24] proposed to synthesize a 4D RGBD LF from a single 2D RGB image based on estimated 4D ray depth. Flynn et al. [12] mapped input views to a set of depth planes of the same perspective through homography transform and then fused them together through two parallel CNNs for learning weights to average the color of each plane. Zhou et al. [25] and Mildenhall and Ben [26] trained a network that inferred alpha and multiplane images. Although utilizing depth information makes it easier to handle inputs with large disparities, most methods cannot achieve acceptable performance for large-baseline sampling, Jin et al. [27] focus on the angular superresolution of light field images with a large-baseline and propose an end-to-end trainable method, by making full use of the intrinsic geometry information of light fields. However, unfortunately inaccurate depth estimation usually happens in challenging scenes that contain significant depth variations, complex lighting conditions, occlusions, non-Lambertian surfaces, etc.

2.2. Depth-Independent LF Reconstruction. Depth-independent LF reconstruction approaches can be considered as an angular dimension upsampling without any geometry information of the scene. Zhouchen and Heung-Yeung [28] proved that for a LF which disparity between neighboring views is less than one pixel, and novel views can be produced using linear interpolation. Some methods have investigated LF reconstruction with specific sampling patterns. Levin and Durand [29] exploited dimensionality gap priors to synthesize novel views from a set of images sampled in a circular pattern. Lixin et al. [14] sampled only the boundary viewpoints or diagonal viewpoints to recover the full LF using sparsity analysis in the Fourier domain. These methods are far from practical application due to the difficulty in capturing input views in specific mode.

Recently, deep learning has been applied in many fields, such as 3D object [30] detection and object recognition [31]. Some learning-based approaches were also proposed for depth-independent reconstruction. Yoon et al. [32] proposed a deep learning framework, in which two adjacent views are employed to generate the interview, while it can only generate novel views at 2x upsampling factor. Wu et al. [19] proposed a blur-restoration-deblur framework as learning-based angular detail restoration on 2D EPIs. Wu et al. [33] further discussed the trade-off either aliasing or blurring problem and designed a Laplacian pyramid EPI (LapEPI) structure that contains both low spatial scale EPI (for aliasing) and high-frequency residuals (for blurring) to solve the trade-off problem. However, the potential of the full LF data is under used in both works. Wang et al. [24] and their subsequent work [35] introduced an end-to-end learning-based pseudo 4DCNN framework using 3D LF volumes with a fixed interpolating rate. On the basis of the pseudo 4DCNN network, Ran et al. [20] added an EPI repair module that magnifies the EPI with arbitrary scale factor.

3. The Proposed Approach

3.1. 4D LF Representation. A 4D LF is usually denoted as $L(u, v, s, t)$, which uses the intersections of light rays with two parallel planes to record light rays, called the two-plane representation (see Figure 1). Each light ray travels from the spatial coordinates (u, v) on the focal plane then to the angular coordinates (s, t) on the camera plane. Therefore, a 4D LF is regarded as a 2D image array with $Cols \times Rows$ images sampled on a 2D angular grid, of which each image is of spatial resolution $W \times H$.

As shown in Figure 1, by fixing u in spatial domain and s in angular domain (or v and t), we can get an EPI map denoted as $E_{u_0,s_0}(v, t)$ (or $E_{v_0,t_0}(u, s)$), which is a 2D slice of the 4D LF. A 3D volume $V_{t_0}(u, v, s)$ (or $V_{s_0}(u, v, t)$) can be produced if we stack EPIs from a column (or a row) views by fixing $t = t_0$ (or $s = s_0$). In the EPI, relative motion between the camera and object points manifests as lines with depth depending slopes. Thus, EPIs can be regarded as an implicit representation of the scene geometry. EPI has a highly clear structure compared with conventional photo images.

The goal of LF angular superresolution is to recover a densely-sampled LF with a resolution of $W \times H \times Cols \times Ro$

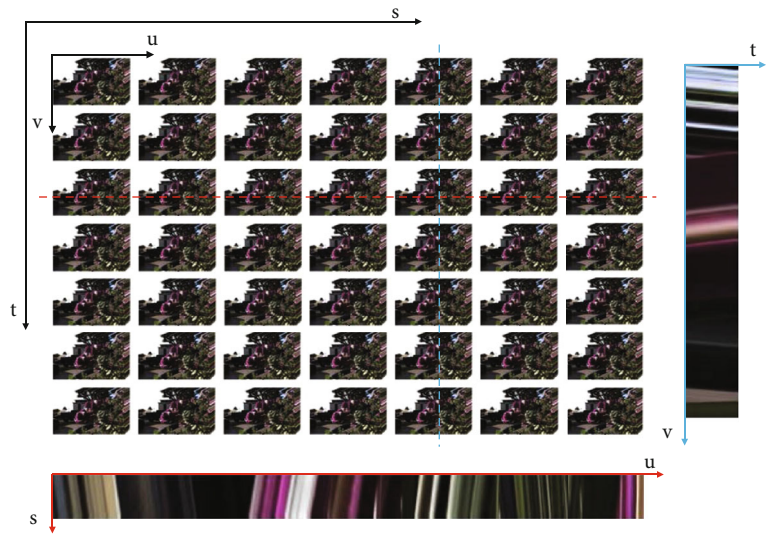


FIGURE 1: A 4D light field $L(u, v, s, t)$ visualization. The horizontal EPI is a 2D (u, s) slice $L(u, v_0, s, t_0)$ by positioning $v = v_0$ and $t = t_0$ (highlighted in red) and the vertical EPI (v, t) by positioning $u = u_0$ and $s = s_0$ (highlighted in blue).

ws from a sparse one with a resolution of $W \times H \times cols \times rows$. The presented framework will be comprehensively introduced in the next section.

3.2. Overview of the Proposed Method. In order to make full use of 4D LF data while circumventing high computational burden, we propose to first synthesize novel views in each row and then in each column, forming a pseudo 4DCNN. Specifically, as shown in Figure 2, for each row/column of views, our network consists of two parts: coarse EPI angular superresolution module based on 2D metalearning and EPI stack refinement module using 3DCNNs. First, we perform these two parts on the 3D row EPI volumes, after by converting the angular from row to column. Then, perform the same operation on the 3D column EPI volumes.

Given an input sparse LF $LF_{\text{in}}(u, v, s, t)$, with the size of $W \times H \times cols \times rows$. First we fix the angular coordinate $s = s_0, s_0 \in \{1, 2, \cdots, \text{cols}\}$ to get the 3D row EPI volumes with the size of $W \times H \times \text{rows}$. Then, we fix the spacial axis $u = u_0, u_0 \in \{1, 2, \cdots, W\}$ and perform the angular superresolution metalearning based network $F(\cdot)$ on each EPI $E_{u_0,s_0}(v, t)$ to get $E^*_{u_0,s_0}(v, t)$ with the size of $H \times \text{Rows}$:

$$E^*_{u_0,s_0}(v, t) = F\left(E_{u_0,s_0}(v, t), r\right), \tag{1}$$

where $r = (Rows - 1)/(rows - 1)$ is the magnification scale factor. The network $N_c(\cdot)$ is followed to recover the high-frequency details of $E^*_{u_0,s_0}(v, t)$ and obtain the LF_{inter} (u, v, s_0, t) with the size of $(W, H, \text{cols}, \text{Rows})$:

$$LF_{\text{inter}}(u, v, s_0, t) = N_c\left(E^*_{s_0}(u, v, t)\right). \tag{2}$$

The angular conversion is performed on $LF_{\text{inter}}(u, v, s_0, t)$, that is, the angular dimension is changed from t to s. First, by fixing the angular coordinate $t = t_0, t_0 \in \{1, 2, \cdots, \text{Rows}\}$, we get the 3D column EPI volumes with the size of $W \times H \times \text{cols}$; then, we extract column EPI by fixing the spacial axis $v = v_0, v_0 \in \{1, 2, \cdots, H\}$ and perform the angular superresolution network $F(\cdot)$ on to obtain superresolved $E^*_{v_0,t_0}(u, s)$, with the size of $W \times \text{Cols}$:

$$E^*_{v_0,t_0}(u, s) = F\left(E_{v_0,t_0}(u, s), r\right). \tag{3}$$

Finally, we use the network $N_r(\cdot)$ to recover the high-frequency details of $E^*_{v_0,t_0}(u, s)$ and obtain the $LF_{\text{out}}(u, v, s, t)$ with the size of $W \times H \times Cols \times Rows$:

$$LF_{\text{out}}(u, v, s, t) = N_r\left(E^*_{t_0}(u, v, s)\right). \tag{4}$$

We use a learnable architecture for both EPI angular superresolution module and EPI stack refinement module, so that the proposed framework can be trained in an end-to-end strategy.

3.3. Coarse EPI Angular Superresolution Module. As shown in Figure 3, for the angular superresolution of EPI, taking $E_{u_0,s_0}(v, t)$ as an example, the acquisition of $E^*_{u_0,s_0}(v, t)$ can be regarded as image superresolution on angular dimension t. We use a feature extraction module to learn the structure of the EPI and then upsample it to the desired resolution. In order to deal with arbitrary scale factor, we use the metalearning strategy in [36]. Since each pixel on the upsampled EPI is predicted via a local kernel. For different scale factors, both the number of the kernels and the weights of the kernels are different. We take advantage of the metalearning to predict the number of the kernels and the weights of the kernels for each scale factor.

Specifically, a feature learning module is settled to extract features from the low-resolution (LR) EPI; then, the upsample module dynamically predicts the weights of the upsampling filters by taking the magnification scale factor as input and using these weights to generate the high-

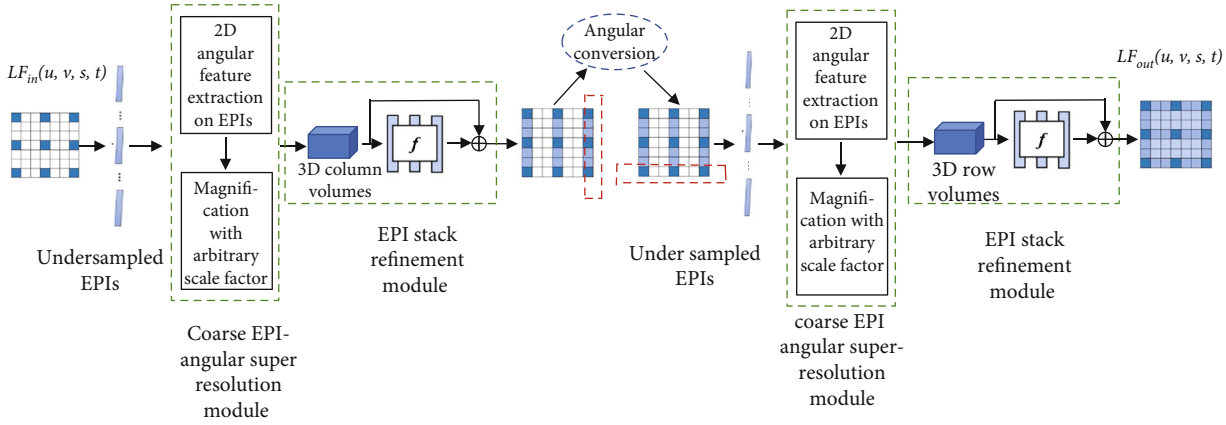


Figure 2: The flowchart of the proposed method for reconstructing a densely-sampled LF from an undersampled LF. Our proposed model consists of two phases, i.e., coarse EPI angular superresolution module and EPI stack refinement module.

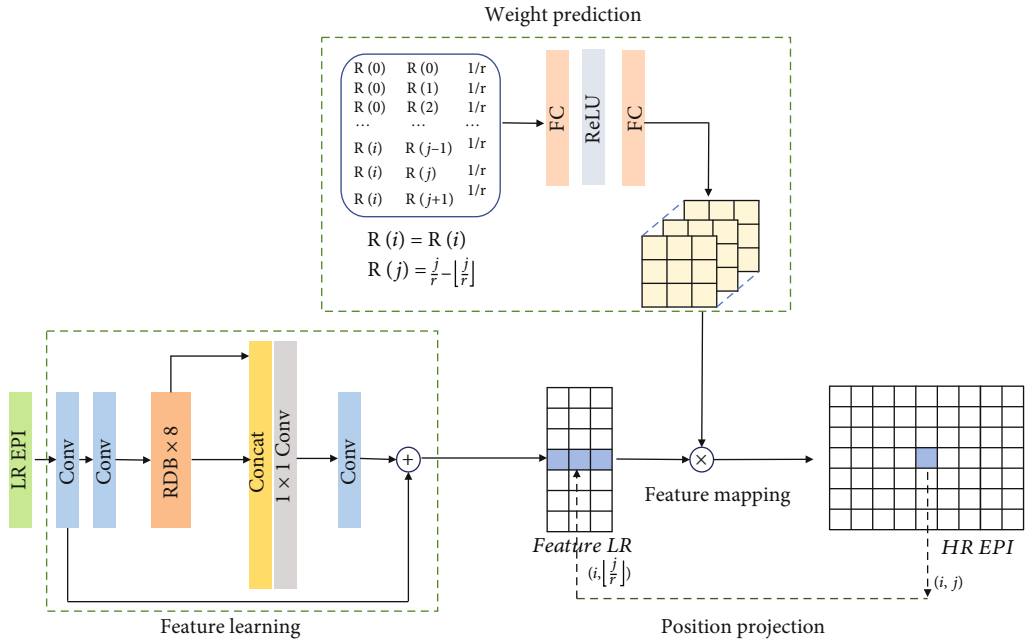


Figure 3: Coarse EPI angular superresolution module structure.

resolution (HR) EPI of arbitrary size. We modify the network to perform magnification only on the angular dimension in EPI.

For the feature learning module, we use the residual dense network (RDN) [37], which is composed of residual learning and dense connections. It consists of 3 layers of 2D CNN and 8 residual dense blocks (RDBs). Each RDB is composed of 8 layers of 2D CNN and ReLU with dense connections.

Taking $E_{u_0,s_0}(v,t)$ as an example, the resolution of the HR EPI is (H, Rows), and the feature is denoted as feature (H, rows). The upsample module can be regarded as the mapping function between EPI $E^*_{u_0,s_0}(v,t)$ with the size of $H \times \text{Rows}$ and feature(H, rows). This part is mainly composed of three steps, namely, position projection, weight prediction, and feature mapping. First, we determine the corresponding pixels across the spatial resolution through the position projection operation, and the specific implementation is as follows:

$$(H, \text{rows}) = T(H, \text{Rows}) = \left(H, \left\lfloor \frac{\text{Rows}}{r} \right\rfloor\right), \tag{5}$$

where T refers to the conversion function, r is the magnification scale factor, and $\lfloor\ \rfloor$ refers to the floor operation. For different magnification scale factors r, we use the weight prediction module to predict the corresponding prediction filter weight:

$$\text{Weight}(H, Rows) = \varphi(V_{H,\text{Rows}}; \theta), \tag{6}$$

$$V_{H,\text{Rows}} = \left(H, \frac{\text{Rows}}{r} - \left\lfloor \frac{\text{Rows}}{r} \right\rfloor\right). \tag{7}$$

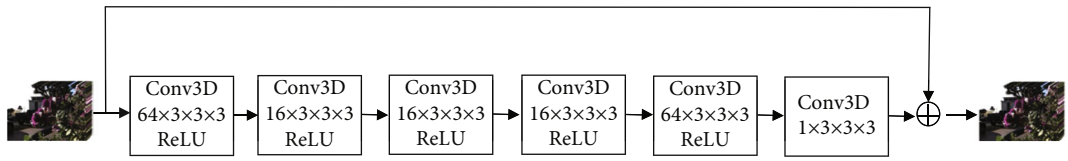


FIGURE 4: Structure of the network for recovering details of 3D volumes. The first 5 layers are followed by a rectified linear unit (ReLU). The final detail restored volume is the sum of the predicted residual and the input.

Among them, $\varphi(\cdot)$ represents the weight prediction network, θ is the network related parameters, and $V_{H,\text{Rows}}$ is the vector of the pixel on the output EPI. Finally, the feature mapping step maps feature(H, rows) with weight(H, Rows) to get the final HR EPI, and matrix multiplication was chosen as its function.

3.4. EPI Stack Refinement Module. The proposed networks $N_c(\cdot)$ and $N_r(\cdot)$ are shown in Figure 4. Both networks consist of an encoder and a decoder, where both of them comprise 3 convolution layers and are exactly symmetric. The first 3D convolutional layer comprises 64 channels with the kernel $3 \times 3 \times 3$, where each kernel operates on 3×3 spatial region across 3 adjacent EPIs. Similarly, the second layer comprises 16 channels with the kernel $3 \times 3 \times 3$. The last layer also comprises 16 channels with the kernel $3 \times 3 \times 3$. Each layer uses a stride of 1 followed by a rectified linear unit (ReLU), i.e., $\sigma(\text{x}) = \max(0, \text{x})$, excepting for the last one. The ultimate output of the network is the sum of the predicted residual and the input 3D EPI volume. 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.

3.5. Perceptual Loss Function. As observed in prior work on image synthesis. Simply comparing the pixel colors of the synthesized image and the reference image could severely penalize perfectly realistic outputs. Instead, we adopt the perceptual loss. The basic idea is to match activations in a visual perception network that is applied to the synthesized image and separately to the reference image.

Let ϕ be a trained visual perception network (we use VGG-16 [38]). Layers in the network represent an image at increasing levels of abstraction: from edges and colors to objects and categories. Matching both lower-layer and higher-layer activations in the perception network guides the synthesis network to learn both fine-grained details and more global part arrangement. Here, g is the image synthesis network being trained and ¦È is the set of parameters of this network.

Let ϕ_1 be a collection of layers in the network ϕ, such that ϕ_1 denotes the input image. Each layer is a three dimensional tensor. For a training pair $(I, L) \in D$, our loss is

$$\iota_{I,L}(\theta) = \sum_l \|\phi_l(I) - \phi_l(g(L;\theta))\|_1. \tag{8}$$

Here, g is the image synthesis network being trained, and θ is the set of parameters of this network. For layers $\phi_l(l \geq 1)$, we use *conv*1_2, *conv*2_2, and *conv*3_2 in VGG-16 [21].

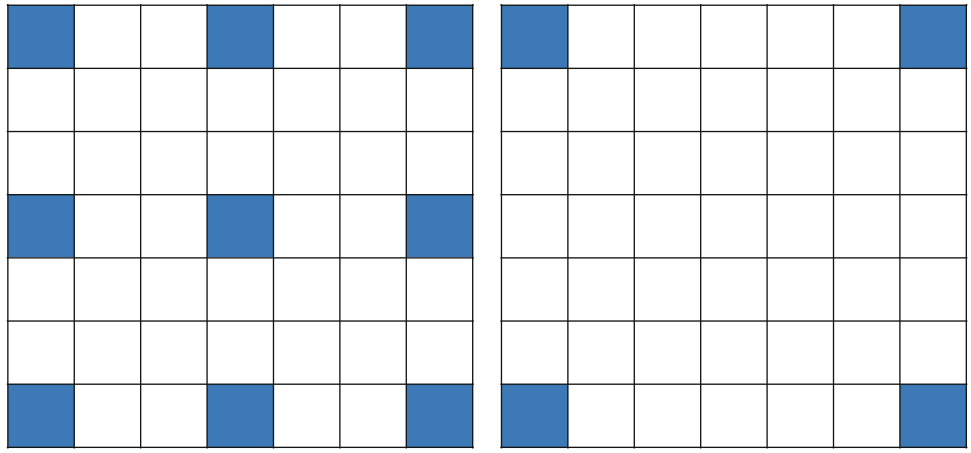
4. Experiments and Results

4.1. Datasets and Training Details. We took real-world LF images captured with a Lytro Illume camera provided by Stanford Lytro LF Archive [39] and Kalantari et al. [17] as well as synthetic LF images from the 4D light field benchmark [40, 41] to train and test the proposed framework. Specifically, 20 synthetic images and 100 real-world images were used for training. 70 LF images captured by Lytro Illum camera were used for real-world scenes test, including 30 test scenes provided by Kalantari et al. [17], 15 LF images from reflective [39] dataset, and 25 LF images from occlusions [39] dataset. 4 LF images from the HCI [41] dataset and 5 LF images from the old HCI [41] dataset were used for synthetic scenes test. We removed the border views and crop the original LF data to 7×7 views as ground truth and then downsampled randomly to 2×2 and 3×3 views as the input. For each LF data, small patches in the same position of each view were extracted to formulate the training LF data. The spatial patch size was 32×32 and the stride was 20.

Similar to other methods, we only processed the luminance Y channel in the YCbCr color space. The framework was implemented with PyTorch. The optimization of end-to-end training was ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the batch size was set to 1. The learning rate was initially set to 10^{-4} and then decreased by a factor of 0.1 every 10 epochs until the validation loss converges. The filters of 3DCNNs were initialized from a zero-mean Gaussian distribution with standard deviation 0.01, and all the bias were initialized to zero.

4.2. Angular Superresolution Evaluation. We used the average value of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) over all synthesized novel views in each scene to quantitatively measure the quality of reconstructed densely-sampled LFs. We designed two experiments to test the angular superresolution quality at different angular resolutions, respectively, reconstructing a 7×7 densely-sampled light field from 2×2 and 3×3 sparse views. Figure 5 demonstrates the sampling patterns.

Our experiments are compared with six state-of-the-art deep learning-based methods designed for densely-sampled light field angular superresolution, namely, Kalantari et al. [17], Wu et al. [18], Wu et al. [19], Wang et al. [35], Jin et al. [27], and Ran et al. [20]. Table 1 shows the properties of the above methods and our proposed method. For the



(a) (b)

Figure 5: Illustration of sampling patterns: (a) 3×3; (b) 2×2.

Table 1: Comparison of properties of different densely-sampled light field angular superresolution algorithms.

Methods	Based on deep learning	Based on geometry	Arbitrary scale factor	One model for various scale factor
Kalantari et al. [17]	√	√	√	−
Wu et al. [18]	√	−	−	−
Wu et al. [19]	√	√	√	−
Wang et al. [35]	√	−	−	−
Jin et al. [27]	√	√	√	−
Li et al. [20]	√	−	√	√
Ours	√	−	√	√

Table 2: Quantitative comparisons (PSNR/SSIM) of the proposed approach with the state-of-the-art ones under task 2×2 to 7×7 on synthetic scenes.

Methods	Kalantari et al. [17]	Wu et al. [18]	Wu et al. [19]	Li et al. [20]	Jin et al. [27]	Ours
HCI	32.85/0.909	26.64/0.744	31.84/0.898	33.14/0.910	34.60/0.937	33.92/0.916
HCI old	38.58/0.944	31.43/0.850	37.61/0.942	38.54/0.944	40.84/0.960	41.50/0.975
Average	36.03/0.928	29.30/0.803	35.05/0.922	36.14/0.928	38.07/0.949	38.13/0.949

compared algorithms, although some of them do support different scale factors [17, 19, 20, 27], they need to train different models for each scale factor separately, while our method is able to perform different scale factors with a single model.

For the task 2×2 to 7×7, we used the 9 synthetic LF images with angular resolution of 9×9, including 4 LF images from the HCI [41] dataset (bicycle bedroom, herbs, and dishes) and 5 LF images from the old HCI [41] dataset (Buddha Buddha2, StillLife, Papillon, and Mona). The central 7×7 views were extracted as ground truth, and 2×2 corner images were taken as input. We carried out comparison with the methods by Wu et al. [19], Kalantari et al. [17], Ran et al. [20], and Wanner and Goldluecke [23]. Table 2 shows the quantitative evaluation of the proposed approach on the synthetic dataset compared with the above methods.

On the two datasets, our proposed method provides an average angular superresolution advantage of 0.06 dB in terms of PSNR. On the old HCI dataset, our proposed method provides an average angular superresolution advantage of 0.66 dB in terms of PSNR and an advantage of 0.015 in terms of SSIM. On the HCI dataset, our method is inferior to Wanner and Goldluecke [23] since they used depth information to connect correspondence in views, but it is still better than other depth-dependent methods such as Kalantari et al. [17] and Wu et al. [19]. This is the cost for depth-free reconstruction. In general, our method has an absolute advantage in non-Lambertian scenes with smaller sampling baseline (the old HCI dataset), while it is also competitive on scenes with large sampling baseline (the HCI dataset). Therefore, we can say that our network has an advantage under task 2×2 to 7×7. We also provided visual comparisons of different methods, as shown in Figure 6. In contrast,

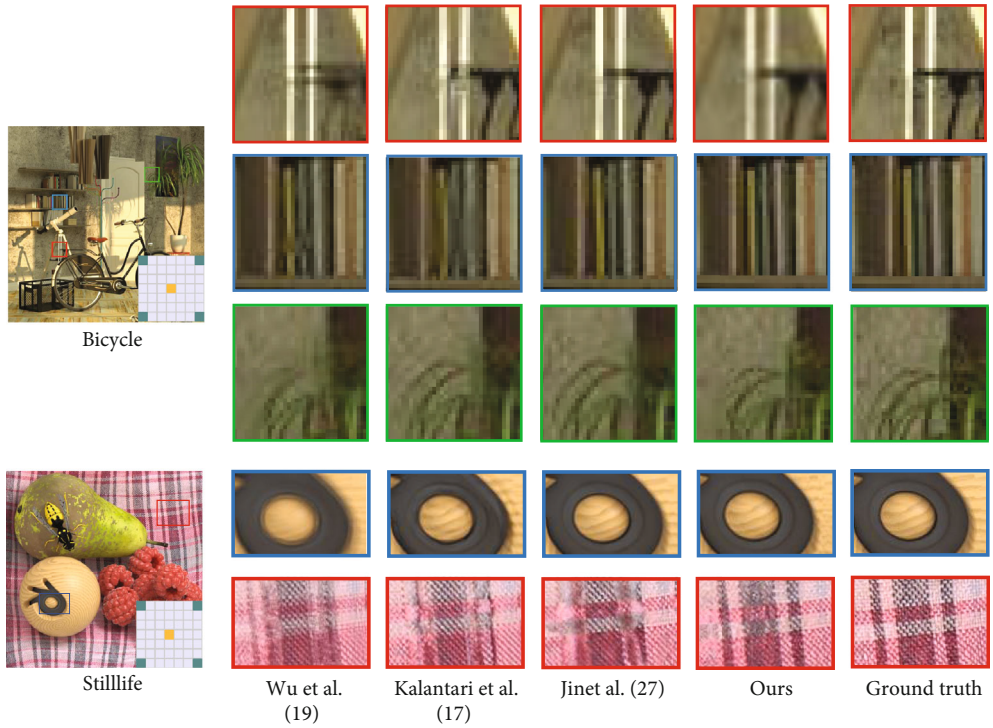


Figure 6: Visual comparison between Wu et al. [19], Kalantari et al. [17], Jin et al. [27], and our approach in synthetic scenes.

Table 3: Quantitative comparisons (PSNR/SSIM) of the proposed approach with the state-of-the-art ones under task 3×3 to 7×7 on real-world scenes.

Methods	Wu et al. [18]	Wang et al. [35]	Li et al. [20]	Ours
30scenes	41.02/0.988	43.82/*0.993*	43.83/0.993	*44.07*/0.988
Occlusions	39.80/0.981	41.23/0.983	41.89/0.984	*42.42/0.988*
Reflective	40.53/0.984	42.33/0.985	42.34/0.982	*42.36/0.986*
Average	40.48/0.985	42.58/*0.988*	42.82/0.987	*43.11/0.988*

our method produces high-quality images which are closer to the ground truth ones.

For the task 3×3 to 7×7, we used the real-world scene LF images with angular resolution of 7×7. We carried out comparison with the method by Wu et al. [18], Wang et al. [35], and Ran et al. [20].

As shown in Table 3, our proposed method performs better for all datasets than comparing methods: with 2.63 dB angular superresolution advantage over Wu et al. [18], 0.53 dB over Wang et al. [35], and 0.41 dB over Ran et al. [20] in terms of PSNR. Experimental results have further proven the advantages of our method.

4.3. Ablation Study. We conducted ablation experiments in terms of network structure. First, in order to verify the effectiveness of the two modules from our coarse-to-fine framework, we removed the EPI stack refinement module and tested only the coarse EPI angular superresolution module. Then, in order to maximize the reconstruction quality and find the optimal structure of the EPI stack refinement module, we evaluated three different structure solutions and selected the best one as the one used in this paper. The specific structure is shown in Figure 7. Scheme 1 is composed of 3 3D convolutional layers, the filter sizes are $5 \times 5 \times 3$, $1 \times 1 \times 3$, and $9 \times 9 \times 3$, and the numbers of output channels are 64, 32, and 1. Each convolutional layer is followed by a ReLU, excepting for the last layer. The filter bank learns the residuals, and the output of the last layer is added to the input as the final result. In general, scheme 1 is used as a residual block, and two residual blocks are connected to form a detail recovery module. The second scheme is an extension of scheme 1 and also refers to the idea of residual network. Scheme 2 is composed of 6 3D convolutional layers and the filter size is set to $3 \times 3 \times 3$. The third one is the method used in this paper. Based on the three schemes, the angular superresolution performance was tested. The specific experimental design was to combine the coarse EPI angular superresolution module with the three schemes.

The experiment was carried out on the HCI dataset, and the task is 2×2 to 7×7. Table 4 shows the experimental results. It can be seen that the angular superresolution quality of the EPI coarse angular superresolution network itself is on the high side. And the addition of the EPI stack refinement module of any scheme can increase the angular

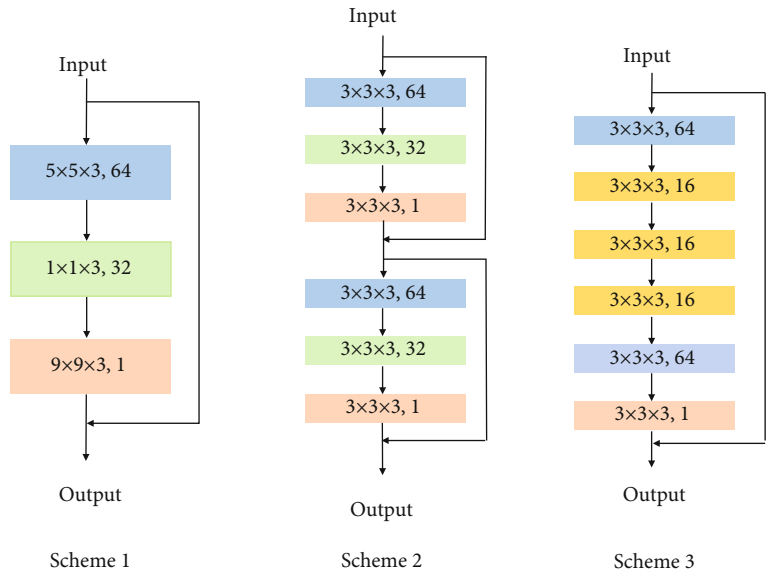


Figure 7: Three different schemes of EPI stack refinement module.

Table 4: Quantitative comparisons (PSNR/SSIM) on synthetic scenes with different EPI stack refinement module under task 2×2 to 7×7.

Methods	HCI
Only 2D EPI angular superresolution module	31.15/0.871
Coarse EPI angular superresolution module + solution 1	31.64/0.902
Coarse EPI angular superresolution module + solution 2	31.96/0.911
Coarse EPI angular superresolution module + solution 3 (ours)	33.92/0.916

superresolution performence on the HCI datasets. This shows that both the coarse EPI angular superresolution module and the EPI stack refinement module play a positive role in improving the quality of the reconstructed light field, and the design of the entire frame structure is reasonable and effective. Between these two modules, the main function of the coarse EPI angular superresolution module is to complete the magnification with any scale factor. The EPI stack refinement module plays an important role in the quality of the reconstructed view, which is responsible for the accurate generation of image textures and complex regions. For the EPI stack refinement module, among the three different structure solutions, scheme 3 has obvious advantages by 1.96 dB.

With the same training environment and training parameters, we used two different loss functions for training and observed the final training results. The first loss function is the edge enhancement loss function proposed in our previous work [20], and the second is the perceptual loss described in Section 3.5. The experiment was carried out on the HCI dataset and the old HCI dataset under the task 2×2 to 7×7. Table 5 shows the experimental results.

Table 5: Quantitative comparisons (PSNR/SSIM) on synthetic scenes with different loss functions under task 2×2 to 7×7.

Methods	Edge enhancement loss function	VGG-16 perception loss
HCI	33.31/0.910	33.92/0.916
HCI old	41.14/0.971	41.50/0.975
Average	37.66/0.944	38.13/0.949

It can be seen that training with perceptual loss has an advantage of 0.47 dB in terms of PSNR and 0.005 in terms of SSIM, compared with training with edge enhancement loss function. We believe that this is because the perceptual loss uses the features extracted by CNN as the object of the loss function, making the global structure and high-level semantics of the synthetic image and the ground truth closer. Compared with edge enhancement loss function, which pays more attention to texture information, perceptual loss performs better on angular superresolution tasks.

5. Conclusion

In this paper, an end-to-end coarse-to-fine framework is proposed to directly synthesize novel views of 4D densely-sampled LF from sparse input views. We combine magnification with arbitrary scale factor network for coarse EPI angular superresolution and 3DCNNs for EPI stack refinement, working in a coarse-to-fine manner and forming a pseudo 4DCNN which can make full use of 4D LF data while circumventing high computational burden. Our framework first reconstructs an intermediate LF by recovering EPI row volumes and then works on EPI column volumes to synthesize the final densely-sampled LF. In addition, metalearning is utilized to upsample EPIs in the coarse EPI angular superresolution module, which enables magnification with arbitrary scale factor with one model. We conducted ablation

experiments on the network structure and loss function, which proved the feasibility of our proposed network framework and the superiority of the perceived loss. Experimental results show that the proposed framework outperforms other state-of-the-art methods.

Data Availability

The light field image data supporting the findings of this study are from previously reported studies and datasets, which have been cited. The processed data are available from the corresponding author upon request.

Conflicts of Interest

The author(s) declare(s) that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62001432 and 61971383), National Key R&D Program of China (Grant No. SQ2020YFF0426386), and the Fundamental Research Funds for the Central Universities (Grant Nos. CUC19ZD006 and CUC21GZ007).

References

[1] L. Marc and H. Pat, "Light field rendering," *ACM*, pp. 31–42, 1996.

[2] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Principles of light field imaging: briefly revisiting 25 years of research," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 59–69, 2016.

[3] J. Fiss, B. Curless, and R. Szeliski, "Refocusing plenoptic images using depth-adaptive splatting," in *IEEE International Conference on Computational Photography*, pp. 1–9, Santa Clara, CA, USA, 2014.

[4] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *Association for Computing Machinery Transactions on Graphics*, vol. 32, no. 4, pp. 1–12, 2013.

[5] J. Chen, J. Hou, Y. Ni, and L. Chau, "Accurate light field depth estimation with superpixel regularization over partially occluded regions," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 27, no. 10, pp. 4889–4900, 2018.

[6] H. Fu-Chung, C. Kevin, and W. Gordon, "The light field stereoscope," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 1–12, 2015.

[7] J. Chai, X. Tong, S. Chan, and S. Shum, "Plenoptic sampling," *SIGGRAPH conference*, pp. 307–318, 2000.

[8] Raytrix, "3d light field camera technology," http://www.raytrix.de/.

[9] T. Bishop and P. Favaro, "The light field camera: extended depth of field, aliasing, and superresolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 972–986, 2012.

[10] V. Boominathan, K. Mitra, and A. Veeraraghavan, *Improving Resolution and Depth-of-Field of Light Field Cameras Using a Hybrid Imaging System*, IEEE International Conference on Computational Photography, 2014.

[11] Penner and Eric, "Soft 3d reconstruction for view synthesis," vol. 36, Tech. Rep. 6, Association for Computing Machinery Transactions on Graphics, 2017.

[12] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, *Deep Stereo: Learning to Predict New Views from the World's Imagery*, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[13] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," vol. 32, Tech. Rep. 4, Association for Computing Machinery Transactions on Graphics, 2013.

[14] S. Lixin, H. Haitham, A. Davis, K. Dina, and D. Fredo, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM Transactions on Graphics*, vol. 34, no. 1, pp. 1–13, 2014.

[15] 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.

[16] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Accelerated shearlet-domain light field reconstruction," *IEEE Journal on Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1082–1091, 2017.

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

[18] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, *Light field reconstruction using deep convolutional network on epi*, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[19] 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.

[20] R. Li, L. Fang, L. Ye, W. Zhong, and Q. Zhang, *Light field reconstruction with arbitrary angular resolution using a deep Coarse-To-Fine framework*, International Forum on Digital TV and Wireless Multimedia Communications, 2021.

[21] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large scale image recognition*, International Conference on Learning Representati, 2014.

[22] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *Association for Computing Machinery Transactions on Graphics*, vol. 32, no. 3, pp. 1–12, 2013.

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

[24] P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, *Learning to Synthesize a 4d Rgbd Light Field from a Single Image*, IEEE Computer Society, 2017.

[25] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–12, 2018.

[26] Mildenhall and Ben, "Local light field fusion," *Transactions on Graphics*, vol. 38, no. 4, pp. 1–14, 2019.

[27] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network,"

Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 11141–11148, 2020.

[28] L. Zhouchen and S. Heung-Yeung, "A geometric analysis of light field rendering," *International Journal of Computer Vision*, vol. 58, no. 2, pp. 121–138, 2004.

[29] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 1831–1838, 2010.

[30] M. Yan, Z. Li, X. Yu, and C. Jin, "An End-to-End Deep Learning Network for 3D Object Detection From RGB-D Data Based on Hough Voting," *IEEE Access*, vol. 8, pp. 138810–138822, 2020.

[31] S. Yang, J. Wang, S. Arif, M. Jia, and S. Zhong, *SAL-Net: Self-supervised attribute learning for object recognition and segmentation*, Wireless Communications and Mobile Computing, 2021.

[32] Y. Yoon, H. G. Jeon, D. Yoo, J. Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 848–852, 2017.

[33] G. Wu, Y. Liu, F. Lu, and T. Chai, "Lapepi-Net: a Laplacian pyramid epi structure for learning-based dense light field reconstruction," https://arxiv.org/abs/1902.06221, 2019.

[34] Y. Wang, L. Fei, Z. Wang, G. Hou, Z. Sun, and T. Tan, *End-to-End view synthesis for light field imaging with pseudo 4dcnn*, European Conference on Computer Vision, 2018.

[35] Y. Wang, F. Liu, K. Zhang, Z. Wang, Z. Sun, and T. Tan, "High-fidelity view synthesis for light field imaging with extended pseudo 4dcnn," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 830–842, 2020.

[36] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, *Meta-Sr: A Magnification-Arbitrary Network for Super-Resolution*, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[37] Y. Zhang, Y. Tian, and Y. Kong, *Residual Dense Network for Image Super-Resolution*, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[38] Q. Chen and V. Koltun, *Photographic Image Synthesis with Cascaded Refinement Networks*, IEEE International Conference on Computer Vision, 2017.

[39] A. S. Raj, "Stanford lytro light field archive," http://lightfields.stanford.edu/.

[40] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, *A Dataset and Evaluation Methodology for Depth Estimation on 4d Light Fields*, Asian Conference on Computer Vision Springer, 2016.

[41] S. Wanner, *Datasets and Benchmarks for Densely Sampled 4d Light Fields*, The Eurographics Association, 2013.