IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 February 2023, accepted 19 March 2023, date of publication 27 March 2023, date of current version 30 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3261967

RESEARCH ARTICLE

High-Fidelity Light Field Reconstruction Method Using View-Selective Angular Feature Extraction

SHUBO ZHOU[1], XUE-QIN JIANG[1], XIAOMING DING[2], AND RONG HUANG[1]
[1]Institute of Information Science and Technology, Donghua University, Shanghai 201620, China
[2]College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

Corresponding author: Rong Huang (rong.huang@dhu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61803372, Grant 62001328, and Grant 62001099; and in part by the Fundamental Research Funds for the Central Universities under Grant 2232021D-34.

ABSTRACT Deep learning (DL) provides an effective approach for light field (LF) reconstruction that aims to synthesize novel views from sparsely-sampled views. However, it is challenging to address domain asymmetry when adopting spatial-angular interaction LF reconstruction methods. To overcome this problem, a view-selective angular feature extraction block (VS-LFAFE) is proposed to obtain full-resolution angular features that enumerate whole viewpoints in a macropixel. By applying the VS-LFAFE, a novel LF reconstruction method is proposed, consisting of two subblocks: a spatial-angular feature extraction and fusion block, and an angular upsampling block. Experimental results demonstrate the effectiveness of the VS-LFAFE, and validate that the proposed method can achieve superior performance compared with the state-of-the-art methods.

INDEX TERMS Light field reconstruction, light field imaging, view-selective angular feature, convolutional neural network.

I. INTRODUCTION

Microlens -based light field (LF) camera [1], [2], [3], [4] can simultaneously capture intensity and angular information of a scene with a microlens array (MLA) between the main optics system and sensor. The additional angular information enables a variety of applications such as refocusing [5], [6], deocclusion [7], [8], depth perception [9], [10], [11], spectral sensing [12], [13], semantic segmentation [14], [15], [16] and reflection removal [17]. However, due to the limitation of sensor resolution, the trade-off between the LF spatial resolution and angular resolution restricts these applications. Currently, two types of methods can mitigate this trade-off: LF spatial super resolution reconstruction [18], [19], [20] and LF reconstruction [21], [22]. This paper explores a novel LF reconstruction method that aims to obtain a densely-sampled LF image from a set of sparsely-sampled sub-aperture images (SAIs).

LF reconstruction, also named as LF angular super resolution [23] or LF view synthesis [24], can be categorized into depth-dependent methods and depth-independent methods. Depth-dependent methods involves two subprocesses: disparity-based view warping and view blending. In these methods, the input LF SAIs are first warped based on the pixelwise disparities to formulate novel angular views. Then, optimization strategies are used to blend and refine the formulated views. On the other hand, depth-independent methods are implemented by extracting and fusing the LF structural features, such as spatial features extracted from the SAIs, angular features extracted from the macropixel image (MacPI), and epipolar plane image (EPI) features extracted from the SAI array. Because the computational efficiency of depth-dependent methods is limited due to disparity estimation and SAI warping processes, this paper focuses on depth-independent methods.

LF reconstruction methods that utilize spatial-angular interaction features [25], [26] have achieved promising results in recent years. However, among these methods, the size of the extracted angular features is much smaller than that of the extracted spatial features, which leads to spatial-angular feature domain asymmetry [27]. To address this issue, upsampling strategies, such as

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

transposed convolution or linear interpolation are required to resize the angular features, which introduces additional error to the interaction spatial-angular features.

To overcome this shortcoming, a novel LF reconstruction method is proposed in this paper. The main contributions are summarized as follows:

- A view-selective angular feature extraction block (VS-LFAFE) is proposed by applying 2×2 convolutional layers to the differently-sampled LF MacPIs. By concatenating the extracted angular features along the channel dimension, a full-resolution angular feature can be obtained by using the pixel shuffling strategy. Based on the ablation results, the extracted full-resolution angular feature is effective in LF reconstruction.
- A CNN-based LF reconstruction network is designed with two subblocks: the spatial-angular feature extraction and fusion block (SA-FEFB) and the angular upsampling block. The SA-FEFB aims to extract discriminative spatial-angular interaction features, and the angular upsampling block aims to angularly upsample the extracted LF features and obtain the reconstructed LF images.
- Extensive experiments on synthetic and real-world LF datasets are conducted to validate the proposed network. Based on the results, the proposed method outperforms other state-of-the-art methods and can preserve accurate parallax structures with reasonable computational efficiency.

The reminder of this paper is organized as follows. In Section II, the related works are reviewed. In Section III, the network framework is proposed. In Section IV, the experimental results are presented. In Section V, conclusions and recommendations for the future work are discussed.

II. RELATED WORKS

A. DEPTH-DEPENDENT LF RECONSTRUCTION

Depth-dependent LF reconstruction methods synthesize novel angular views from a set of sparsely-sampled SAIs with guidance from the disparity estimation results. Georgiev et al. [28] synthesized novel views with a weighted interpolation approach, in which the weight was obtained by computing the flow between the reference SAI and its neighbourhood SAIs. Wanner and Goldluecke [29] proposed a regularization-based LF reconstruction method incorporating subpixel-level disparity maps as priors. As an addition step from the LF reconstruction, the disparity estimation process was formulated as a global optimization problem by using the EPI features. Kalantari et al. [24] proposed a learning-based method for LF reconstruction, which consisted of two subprocesses: disparity estimation and color estimation. Sequential CNNs were used in both subprocesses, and the network was trained by minimizing the error between the synthesized and ground truth (GT) images. Shi et al. [30] proposed a pixel and feature fused method. In this method, a disparity map was obtained from a lightweight optical flow estimation network, and two reconstruction modules were designed in the pixel and feature domains. Jin et al. [31] proposed a CNN-based method in a coarse-to-fine manner. In this method, densely-sampled angular views were first synthesized by using a confidence-based blending strategy, and then, an LF refinement module was used to recover the LF parallax structure. They [23] also proposed an end-to-end learning-based method with two learnable modules and a physically based module. In this method, the depth estimation module was designed to explicitly model scene geometry, the physically based module was designed to warp the angular views, and the light field blending module was designed for light field reconstruction. Meng et al. [32] proposed a learning-based method by jointly modeling the epipolar property and occlusions. In this method, a warping confidence map was developed to handle the occlusions, and 4D CNN was used to refine the synthesized angular views. Guo et al. [33] proposed a CNN-based method for wide-baseline LF reconstruction. In this method, a learnable dynamic interpolation block was proposed to replace the commonly used geometry warping operation, and the weights in dynamic interpolation were learned by a lightweight neural network.

B. DEPTH-INDEPENDENT LF RECONSTRUCTION

The depth-independent LF reconstruction methods extract and fuse the LF structural features to learn the implicit relationships between the densely-sampled and sparsely-sampled LFs. Based on the extracted LF structural features, this kind of methods can be categorized into EPI-based methods and spatial-angular interaction methods.

EPI is generated by fixing one spatial and angular coordinate in 4D light field data. Based on the slope of the EPI lines, the disparity and occlusion patterns can be estimated, and can thus be used for LF reconstruction. Wang et al. [34] proposed a pseudo 4D CNN-based method. In this method, the pseudo 4D CNN was formulated with 2D stride convolutions on stacked EPIs and detail-restoration 3D CNNs connected with angular conversions. Wu et al. [35] proposed a CNN-based method by taking advantage of the pattern that a sheared EPI exhibits a clearer geometric structure. In this method, the CNN was elaborately designed to learn the similarities between the input sheared EPIs and the ground truth EPIs. They [36] also proposed an end-to-end deep anti-aliasing neural network (DA^2N) to solve the challenges of large disparity and non-Lambertian effect. In this method, pseudo EPIs from unstructured LFs were used in the training process. Suhail et al. [37] proposed a two-stage transformer-based model. In this method, the features were first aggregated along the epipolar line dimension and then aggregated along the reference view dimension to produce color information. Yang et al. [38] proposed a 4D convolution-based method in which three paralleled 4D convolutions with residual mechanisms were designed to simultaneously extract EPI features and scene features.

The other branch of depth-independent LF reconstruction methods is the spatial-angular interaction methods, which extract spatial and angular features on an SAI array or

MacPI. Yeung et al. [25] proposed a CNN-based method in a ''coarse-to-fine'' manner, in which spatial-angular alternating convolutions were designed to learn the LF intrinsic spatial-angular features. Meng et al. [39] proposed a high-dimensional convolution-based method, which consisted of a residual network that restores local spatio-angular information and a refinement network that reconstructs the spatial details of the scenes. Hu et al. proposed [40] a CNN-based method in which U-Net [41] was used to extract the hierarchical features, and spatio-angular separable (SAS) convolution layers were used to separate and fuse the spatial and angular features. By applying the spatial-angular alternating mechanism, this method can be trained on larger patches and can improve performance particularly in occluded regions. They [42] also proposed a spatio-angular dense network, in which the correlation blocks were proposed to model the correlation information, and the spatio-angular dense skip connections were proposed to improve the information flow within spatial and angular domains. Cheng et al. [43] proposed a spatial-angular versatile convolution (SAV-conv) module by combining the spatial-angular separable convolution (SAS-conv) and spatial-angular correlated convolution (SAC-conv), which could embed global and robust geometry information into the extracted features.

For spatial-angular interaction methods, 4D convolutions and spatial-angular interaction convolutions are frequently used to extract discriminative features. 4D convolutions can fully extract the high-dimensional features in one convolutional layer, but the computational complexity is relatively high. Therefore, this paper focuses on the LF reconstruction method based on spatial-angular interaction convolutions.

III. METHODOLOGY

A. OVERALL NETWORK

Based on the two-plane model [44], an LF image can be denoted as $\mathcal{I} \in \mathbb{R}^{U\times V\times H\times W}$,where (U, V) are the angular resolutions, and (H, W) are the spatial resolutions. In this paper, we set $U = V = A$, thus a marcropixel is grouped as a square matrix. LF reconstruction aims to reconstruct an LF Image $\mathcal{I}' \in \mathbb{R}^{\alpha U\times\alpha V\times H\times W}$ from $\mathcal{I}$, where α is the upsampling rate.

The overall network framework is illustrated in Fig. 1(a). The input of the network is an angularly sparsely-sampled SAI array, which is first reshaped to a MacPI $\mathcal{I}_m \in \mathbb{R}^{AH\times AW}$ and fed to a 3×3 convolutional layer to obtain the shallow features $\mathcal{F}_{F,0}$ with a size of $\mathbb{R}^{AH\times AW\times C}$, where C is the channel depth. The dilation of the 3×3 convolutional layer is set to A to avoid angular aliasing. Then, $\mathcal{F}_{F,0}$ is fed to the SA-FEFB to obtain spatial-angular interaction features $\mathcal{F}_F$. An SA-FEFB consists of k_0 spatial-angular interacted feature extractors (SA-FEBs). To fully exploit each stage information, features from each SA-FEB are concatenated, and then the concatenated feature is fed to a 1×1 convolutional layer to fuse the channel information. The overall calculation process for the SA-FEFB can be expressed as:

$$\mathcal{F}_F = H_{1\times1}\left(\left[\mathcal{F}_{F,1}, \mathcal{F}_{F,2}, \cdots, \mathcal{F}_{F,k_0}\right]\right), \tag{1}$$

where $H_{1\times1}$ is the 1×1 convolutional layer, $\mathcal{F}_{F,k_0}$ is the feature calculated from the k_0th SA-FEB, and [.] is the concatenation operator along the channel dimension.

Following [45], an angular upsampling module is designed to upsample the fused feature. First an $A \times A$ stride convolutional layer is used to obtain angularly downsampled features with a size of $\mathbb{R}^{H\times W\times C}$, in which the stride is set to A. Then, a 1×1 convolutional layer is used to expand the channel depth to $(\alpha A)^2C$. A 2D pixel shuffling layer is used to produce an upsampled feature $\mathcal{F}_U \in \mathbb{R}^{\alpha AH\times\alpha AW\times C}$. The overall calculation process of the angular upsampling module can be expressed as:

$$\mathcal{F}_U = H_{p\_s\_\alpha A}\left(H_{1\times1}\left(H_{A\times A}\left(\mathcal{F}_F\right)\right)\right), \tag{2}$$

where $H_{p\_s\_\alpha A}$ is the pixel shuffling layer with an upsampling rate of αA, and $H_{A\times A}$ is the $A \times A$ stride convolutional layer.

Finally, a 1×1 convolutional layer is employed to fuse the channels to obtain the LF reconstruction results.

B. SA-FEB

The core block of the proposed network is the SA-FEFB, which consists of several SA-FEBs. The SA-FEB aims to extract and fuse spatial and angular features from the MacPI-sampled shallow feature, which involves three sub-blocks: LF spatial feature extraction block (LF-SFE), view-selective LF angular feature extraction block (VS-LFAFE) and spatial-angular feature fusion block.

1) LF-SFE

The LF-SFE aims to extract each SAI's spatial feature. This block consists of two feature extraction groups, where each group involves k_1 spatial feature extractors (SFEs), as illustrated in Fig. 1(b). The SFE is designed with a ''3×3 convolution-PRelu-3×3 convolution'' structure, where the dilation of the 3×3 convolutional layer is set to A. To fully exploit each stage information in a spatial feature extraction group, the feature from each SFE is concatenated and fused with a 1×1 convolutional layer. Finally, the extracted feature is added to the initial shallow feature to form a global residual connection. After processing with the LF-SFE, the size of the extracted spatial feature $\mathcal{F}_s$ is $AH \times AW \times C$.

2) VS-LFAFE

Existing methods always use a single $A \times A$ convolutional layer as an angular feature extractor (Sin-LFAFE), in which the stride is set to A. As illustrated in Fig. 2, this kind methods produce domain asymmetry between the extracted spatial and angular features. Also, the reference pixel of the $A \times A$ convolutional layer is always located at a specific position (left corner pixel) in the macropixel region, thus, the convolutional layer can only extract single pattern angular features.

To overcome these limitations, VS-LFAFE is proposed to extract full-resolution angular features by enumerating all

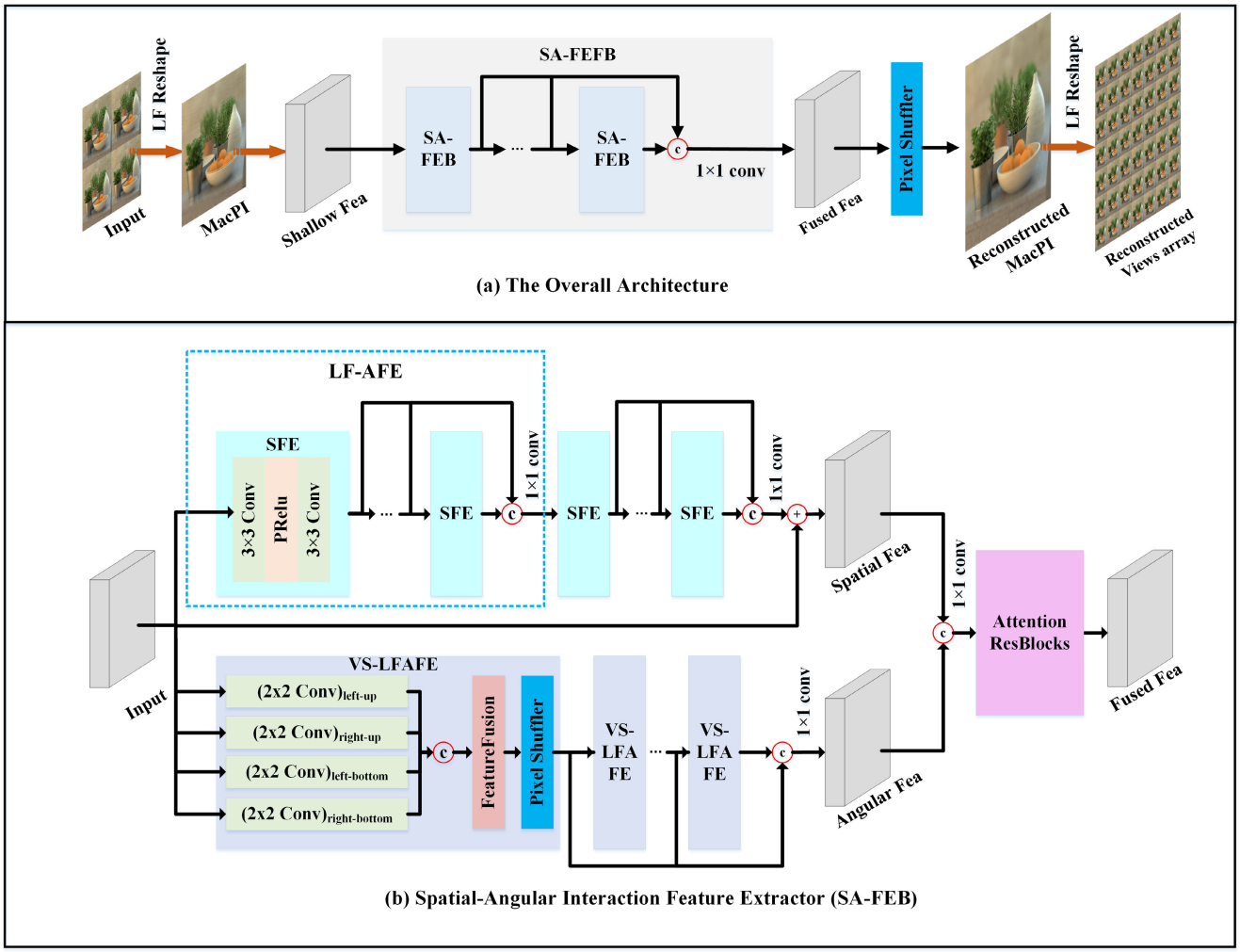

FIGURE 1. Proposed network framework.

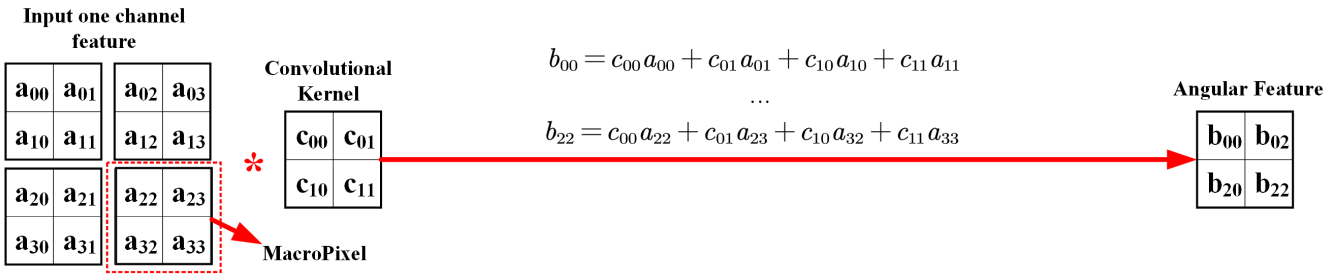

FIGURE 2. The single-pattern LF angular feature extractor (Sin-LFAFE) is implemented by using one stride convolutional layer. *a* is the coefficients of the input one channel feature map. **c** is the learnable convolutional weights, and **b** is the coefficients of the extracted angular feature.

pixels in a macropixel. As illustrated in Fig. 3, the VS-LFAFE consists of four kinds of 2×2 convolutional layers: the left-up, left-bottom, left-right and right-bottom convolutional layer. For each 2×2 convolutional layer, the stride is also set to A.

The left-up convolutional layer is the same as the Sin-LFAFE, which enumerates left-top $(A-1)\times(A-1)$ pixels in each macropixel to obtain corresponding angular features. Assume that $\mathcal{F}_{F,0,0}$ is the first macropixel of the shallow feature $\mathcal{F}_{F,0}$,

$$\mathcal{F}_{F,0,0} = \mathcal{F}_{F,0}(1:A, 1:A). \tag{3}$$

For pixel $\mathcal{F}_{F,0,0}(i_0, j_0)$ where i_0 and j_0 range from 1 to $A-1$, the corresponding left-up convolutional layer can be expressed as:

$$\mathcal{F}_{L,U,i_0,j_0} = H_{2\times2}\left(\mathcal{F}_{F,0}(i_0:AH, j_0:AW)\right), \tag{4}$$

FIGURE 3. The proposed VS-LFAFE is implemented using four 2 × 2 stride convolutional layers.

where $\mathcal{F}_{L,U}$ is the extracted angular feature, and $H_{2\times2}$ is the 2×2 stride convolutional layer. To make the reference pixel in $\mathcal{F}_{F,0,0}$ at (i_0, j_0), the first $i_0 - 1$ rows and $j_0 - 1$ columns of the input feature map are cropped.

The left-bottom convolutional layer aims to extract angular features with the reference pixel in $\mathcal{F}_{F,0,0}$ at (A, j_1), where j_1 ranges from 1 to $A - 1$, which can be expressed as:

$$\mathcal{F}_{L,B,j_1} = F_{ud}\left(H_{2\times2}\left(F_{ud}\left(\mathcal{F}_{F,0}(1:AH, j_1:AW)\right),1\right),1\right), \tag{5}$$

where $\mathcal{F}_{L,B}$ is the extracted angular feature, and F_{ud} is the up-down flipping operation. Two up-down flipping operations are conducted to align $\mathcal{F}_{L,B}$ with $\mathcal{F}_{L,U}$.

The left-right convolutional layer aims to extract angular features with the reference pixel in $\mathcal{F}_{F,0,0}$ at (i_1, A), where i_1 ranges from 1 to $A - 1$, which can be expressed as:

$$\mathcal{F}_{L,R,i_1} = F_{lr}\left(H_{2\times2}\left(F_{lr}\left(\mathcal{F}_{F,0}(i_1:AH, 1:AW)\right),1\right),1\right), \tag{6}$$

where $\mathcal{F}_{L,R}$ is the extracted angular feature, and F_{lr} is the left-right flipping operation.

The right-bottom convolutional layer aims to extract angular features with the reference pixel at $\mathcal{F}_{F,0,0}(A, A)$, which can be expressed as:

$$\mathcal{F}_{R,B} = F_{ot}\left(H_{2\times2}\left(F_{ot}\left(\mathcal{F}_{F,0}\right),1\right),1\right), \tag{7}$$

where $\mathcal{F}_{R,B}$ is the extracted angular feature, and F_{ot} is the origin transformation operation.

By concatenating the extracted angular feature along the channel dimension, a channel-expanded feature $\mathcal{F}_{F,C}$ can be obtained with a size of $H \times W \times A^2C$. Then, a 2D pixel shuffling layer is used to produce an upsampled feature with a size

TABLE 1. Comparative ablation results evaluated by PSNR.

	SFE	Sin-LFAFE	VS-LFAFE	Attention ResNet	30scene	Occlusion	Reflective	Average
$model1$	✓			✓	42.63	38.59	38.39	39.87
$model2$			✓	✓	39.82	37.40	37.73	38.32
$model3$	✓		✓		43.28	39.07	38.93	40.43
$model4$	✓	✓		✓	42.86	38.58	38.65	40.03
Proposed Network	✓		✓	✓	43.74	39.31	39.14	40.73

of $AH \times AW \times C$. Finally, by cascading k_2 VS-LFAFEs, the final angular feature $\mathcal{F}_a$ can be obtained. Similar to the spatial feature extraction block, the feature from each VS-LFAFE is also concatenated and fused with a 1×1 convolution.

3) SPATIAL-ANGULAR FEATURE FUSION BLOCK

An attention-based residual block is used to fuse more discriminative features from the extracted spatial and angular features. In this block, the spatial and angular features are first concatenated and then fused with a 1×1 convolutional layer. The size of the fused feature is $AH \times AW \times C$. Then, an attention residual block is cascaded, which contains k_3 residual blocks and a channel attention block. The overall spatial-angular feature fusion process can be expressed as:

$$\mathcal{F}_{sa} = H_{att}(H_{1\times1}([\mathcal{F}_s, \mathcal{F}_a])), \tag{8}$$

where $\mathcal{F}_{sa}$ is the spatial-angular fused feature, and H_{att} is the attention residual block.

IV. EXPERIMENTS

A. DATASETS AND IMPLEMENTING DETAILS

Two synthetic datasets (i.e., the HCInew and HCIold datasets) and two real-world datasets (i.e., the 30scene and STFlytro datasets) are used to train and test the proposed network. Following [23], for the synthetic datasets, 20 scenes are used for training, and 4 scenes from the HCInew dataset, and 5 scenes from the HCIold dataset are used for testing. For the real-world datasets, 100 scenes are used for training, and 30 scenes from the 30scene dataset, 25 scenes from the occlusions category and 15 scenes from the reflective category in the STFlytro datasets are used for testing.

Following [23], [45], this paper focuses on reconstructing 7×7 densely-sampled LF data from 2×2 sparsely-sampled LF data. Therefore, during data preparation, ground truth (GT) samples are obtained by angularly cropping the central 7×7 SAIs of each LF. The input samples are generated using the 2×2 corner SAIs of the GT samples. To save GPU memory, each SAI is cropped to patches with 64×64 pixels during the training process. Some data argumentation strategies are performed to enhance the robustness, including horizontal flipping, vertical flipping and 90-degree rotation.

In the training process, C is 64, k_0 is 2, k_1 is 4, k_2 is 8, and k_3 is 2. The proposed network is trained with an L_1 loss and optimized by the Adam optimizer [46] with a batch size of 4. The initial learning rate is set to 2×10^{-4}, and decreased by 0.65 every 10 epochs. The proposed network is implemented in the PyTorch framework with an Nvidia GTX2080Ti GPU, and stopped after approximately 40 epochs for the synthetic dataset, and approximately 70 epochs for the real-world dataset.

In the testing process, each SAI is cropped to patches with a size of 128×128 pixels. In the cropping process, the stride is set to 64 pixels. Then, the SAI patch array is resampled to a MacPI and fed to the pretrained model to obtain the LF reconstruction result. PSNR and SSIM are used to evaluate the Y channel image. Note that only the 45 reconstructed views are evaluated.

B. ABLATION STUDY

In this subsection, some experiments are conducted to investigate the efficiency of the module in the proposed network. By testing each model on the real-world dataset, the ablation results are listed in Table. 1.

1) PROPOSED NETWORK W/O ANGULAR FEATURE

In this part, only the LF spatial feature is used to formulate $model1$. Without utilizing the angular feature, the PSNR results are decreased by 0.86dB. This is because, the angular feature incorporates abundant context information among the SAIs, which is beneficial in synthesizing novel views.

2) PROPOSED NETWORK W/O SPATIAL FEATURE

In this part, only the LF angular feature is used to formulate $model2$. Without utilizing the spatial feature, the PSNR results are decreased by 2.41dB. This is because, the spatial feature incorporates flexible texture and color information, which is important to reconstruct accurate textures in novel view reconstruction process.

3) PROPOSED NETWORK W/O ATTENTION RESBLOCKS

In this part, the attention resblock, which fuses the extracted spatial and angular features, is replaced with a 1×1 convolutional layer to formulate $model3$. Without using the attention Resblocks, the PSNR results are decreased by 0.30dB. Based on the comparative results, extracting more discriminative features from the spatial-angular interaction features is beneficial in reconstructing more accurate novel views.

4) EFFECTIVENESS INVESTIGATION OF THE VS-LFAFE

In this part, the VS-LFAFEs are replaced by the Sin-LFAFEs to formulate $model4$. Although the spatial and angular

TABLE 2. PSNR and SSIM results achieved by different LF reconstruction methods.

	HCIold		HCInew		30scenes		Occlusion		Reflective	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
P4DCNN [34]	29.61	0.819	35.73	0.898	38.22	0.970	35.42	0.962	35.96	0.942
ShearedEPI [35]	31.84	0.898	37.61	0.942	39.17	0.975	34.41	0.955	36.38	0.944
LFASR-geo [23]	34.60	0.937	40.84	0.960	42.53	0.985	38.36	0.977	38.20	0.955
FS-GAF [31]	**37.14**	0.966	41.80	0.974	42.75	0.986	38.51	0.979	38.35	0.957
Yeung et. al. [25]	32.30	0.900	39.69	0.941	42.77	0.986	38.88	0.980	38.33	0.960
Proposed Network	34.63	**0.972**	**42.27**	**0.980**	**43.74**	**0.995**	**39.31**	**0.991**	**39.14**	**0.979**

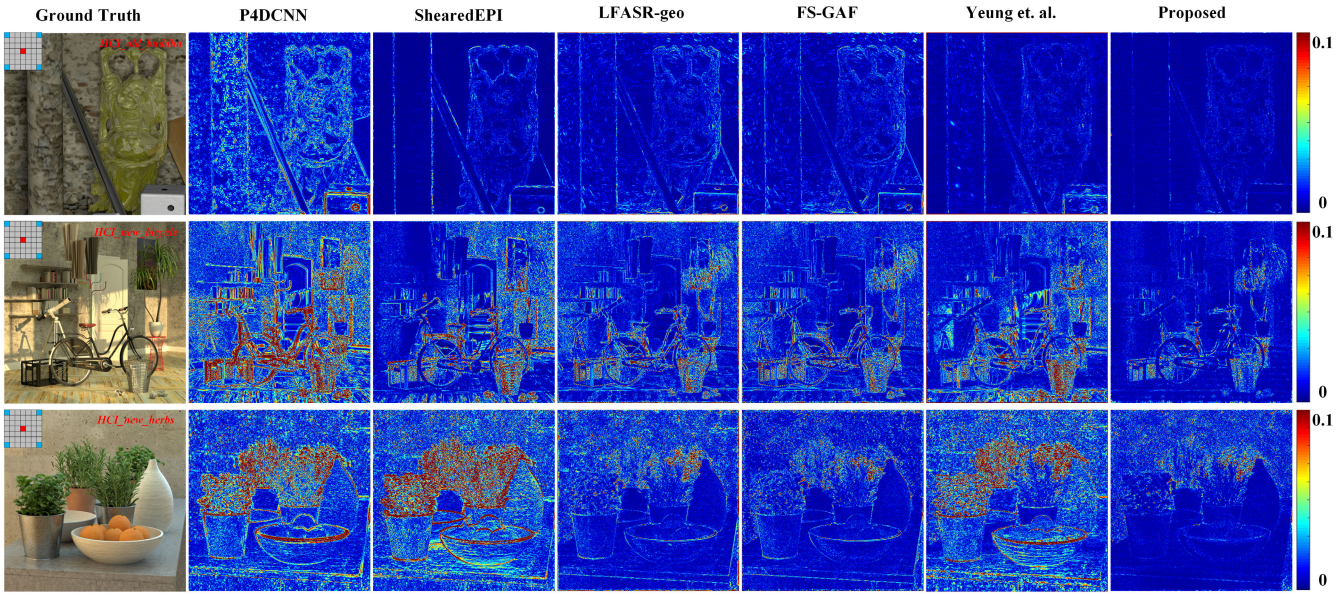

FIGURE 4. Visual comparison results of the synthetic LF images.

features are both used, the PSNR performance of *model*4 is still decreased by 0.70dB. This is because, the proposed VS-LFAFE can extract different pattern angular features and formulate full-resolution angular features without applying upsampling strategies, which makes the module able to obtain more integrated angular features.

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

In this subsection, five state-of-the-art methods are adopted for comparison with the proposed method, which includes P4DCNN [34], ShearedEPI [35], LFASR-geo [23], FS-GAF [31] and Yeung et al. [25]. All of the state-of-the-art methods are retrained to adapt to the "$2 \times 2 \rightarrow 7 \times 7$" LF reconstruction condition.

1) QUANTITATIVE COMPARISON

The PSNR and SSIM results are shown in Table 2. The best results are in bold, and the second-best results are underlined. The P4DCNN and ShearedEPI methods are EPI-based methods. Because only 2 rows or columns of EPI features are used, these methods have difficulties in restoring accurate spatial information, thus, the reconstruction performance is limited. The LFASR-geo and FS-GAF methods are both disparity-dependent methods. Applying disparity estimation and feature warping operations, these methods can achieve better performance in terms of PSNR and SSIM. The method proposed by Yeung et al. is a spatial-angular interaction method, that can also achieve comparable performance. Among all these methods, the proposed network can achieve the best score in all 5 datasets in terms of the SSIM and can achieve the best score in 4 datasets in terms of the PSNR. By extracting view-selective angular features and incorporating them with spatial features, the proposed method can fully utilize the high-dimensional LF features, thus, the novel views can be well reconstructed.

2) QUALITATIVE COMPARISON

Fig. 4 shows the visual compassion results of the synthetic LF images, and Fig. 5 shows the visual compassion results of the real-world LF images. Based on the results of the error map, the EPI methods can not restore the texture features well because the EPI features are too sparse. The two disparity-dependent methods significantly improve the performance, however, the performance in the occluded region

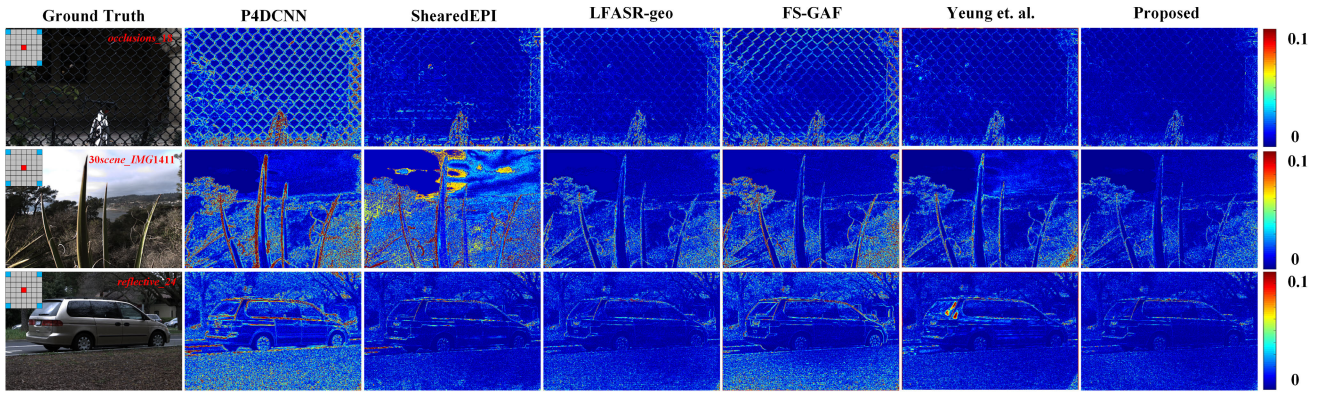

FIGURE 5. Visual comparison results of the real-world LF images.

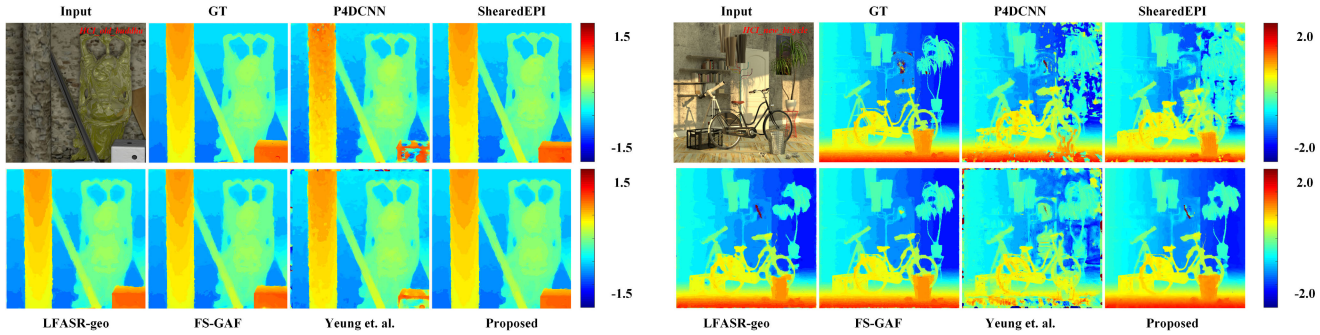

FIGURE 6. Disparity estimation results of the synthetic LF image buddha and bicycle.

is also limited, as in which the disparity estimation accuracy is relatively low. By fully utilizing the spatial and angular features, the proposed method can achieve closer results and produce fewer artifices.

3) ANGULAR CONSISTENCY COMPARISON

The target of LF reconstruction is not only angularly upsampling the sparse sampled LFs but also preserving the parallax structure. Therefore, the disparity estimation method SPO [47] is used to obtain the disparity map to evaluate the angular consistency. Results are shown in Fig. 6, based on which the disparity estimation results of LFASR-geo, FS-GAF and the proposed method can achieve comparable performance compared with the GT result, which indicates that the proposed method can produce results with high angular consistency.

4) COMPUTATIONAL EFFICIENCY ANALYSIS

The inference time is used to evaluate the computational efficiency. The evaluation is performed on the "$2 \times 2 \rightarrow 7 \times 7$" task with an SAI spatial resolution of 512×512 pixels. All methods were evaluated on the same GPU of an NVIDIA GeForce RTX 2080 Ti, and the results are listed in Table 3. Based on the results, the proposed method can achieve the lowest inference time.

TABLE 3. Comparisons of the inference time.

	Inference Time (s)	HCInew	30scene
P4DCNN	1.07	35.73/0.898	38.22/0.970
ShearedEPI	101.70	37.61/0.942	39.17/0.975
LFASR-geo	3.49	40.84/0.960	42.53/0.985
FS-GAF	40.21	41.80/0.974	42.75/0.986
Yeung	0.85	39.69/0.941	42.77/0.986
Proposed Network	**0.50**	**42.27/0.980**	**43.74/0.995**

V. CONCLUSION

In this paper, a novel LF reconstruction method is proposed by designing a CNN-based network interacting the spatial and angular features. In the proposed network, a novel angular feature extractor (VA-LFAFE) is designed with 4 branches of 2×2 convolutional layers to tackle the domain asymmetry between the spatial and angular features. Extensive experiments demonstrate that the proposed framework can achieve state-of-the-art performance with reasonable computational efficiency. In the future, the proposed network can be improved by designing an occlusion-aware network to accurately handle occlusions in LF reconstruction process.

REFERENCES

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1–11, 2005.

[2] X. Lin, J. Wu, G. Zheng, and Q. Dai, "Camera array based light field microscopy," *Biomed. Opt. Exp.*, vol. 6, no. 9, pp. 3179–3189, Sep. 2015.

[3] Y. Zhang, C. Shen, W. Yang, and J. Yu, "Multi-flash light field photography," *IEEE Access*, vol. 7, pp. 52132–52141, 2019.

[4] E. Shafiee and M. G. Martini, "Datasets for the quality assessment of light field imaging: Comparison and future directions," *IEEE Access*, vol. 11, pp. 15014–15029, 2023.

[5] C. Zhang, G. Hou, Z. Zhang, Z. Sun, and T. Tan, "Efficient auto-refocusing for light field camera," *Pattern Recognit.*, vol. 81, pp. 176–189, Sep. 2018.

[6] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective light field refocusing for camera arrays using Bokeh rendering and super-resolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, Jan. 2019.

[7] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, and Y. Guo, "DeOccNet: Learning to see through foreground occlusions in light fields," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 118–127.

[8] X. Wang, J. Liu, S. Chen, and G. Wei, "Effective light field de-occlusion network based on Swin Transformer," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 1, 2022, doi: 10.1109/TCSVT.2022.3226227.

[9] T. Yan, F. Zhang, Y. Mao, H. Yu, X. Qian, and R. W. H. Lau, "Depth estimation from a light field image pair with a generative model," *IEEE Access*, vol. 7, pp. 12768–12778, 2019.

[10] Y. Wang, L. Wang, Z. Liang, J. Yang, W. An, and Y. Guo, "Occlusion-aware cost constructor for light field depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19809–19818.

[11] F. Liu, S. Zhou, Y. Wang, G. Hou, Z. Sun, and T. Tan, "Binocular light-field: Imaging theory and occlusion-robust depth perception application," *IEEE Trans. Image Process.*, vol. 29, pp. 1628–1640, 2020.

[12] Z. Xiong, L. Wang, H. Li, D. Liu, and F. Wu, "Snapshot hyperspectral light field imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3270–3278.

[13] X. Ding, L. Hu, S. Zhou, X. Wang, Y. Li, T. Han, D. Lu, and G. Che, "Snapshot depth–spectral imaging based on image mapping and light field," *EURASIP J. Adv. Signal Process.*, vol. 2023, no. 1, pp. 1–18, 2023.

[14] H. Sheng, R. Cong, D. Yang, R. Chen, S. Wang, and Z. Cui, "UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7880–7893, Nov. 2022.

[15] C. Jia, F. Shi, M. Zhao, Y. Zhang, X. Cheng, M. Wang, and S. Chen, "Semantic segmentation with light field imaging and convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[16] M. Z. Alam, S. Kelouwani, J. Boisclair, and A. A. Amamou, "Learning light fields for improved lane detection," *IEEE Access*, vol. 11, pp. 271–283, 2022.

[17] Y. Ni, J. Chen, and L.-P. Chau, "Reflection removal on single light field capture using focus manipulation," *IEEE Trans. Comput. Imag.*, vol. 4, no. 4, pp. 562–572, Dec. 2018.

[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. Workshop (ICCVW)*, Dec. 2015, pp. 24–32.

[19] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 290–308.

[20] S. Zhou, L. Hu, Y. Wang, Z. Sun, K. Zhang, and X.-Q. Jiang, "AIF-LFNet: All-in-focus light field super-resolution method considering the depth-varying defocus," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jan. 16, 2023, doi: 10.1109/TCSVT.2023.3237593.

[21] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6319–6327.

[22] N. Meng, Z. Ge, T. Zeng, and E. Y. Lam, "LightGAN: A deep generative model for light field reconstruction," *IEEE Access*, vol. 8, pp. 116052–116063, 2020.

[23] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11141–11148.

[24] 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. 1–10, Nov. 2016.

[25] 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. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 137–152.

[26] G. Wu, Y. Wang, Y. Liu, L. Fang, and T. Chai, "Spatial-angular attention network for light field reconstruction," *IEEE Trans. Image Process.*, vol. 30, pp. 8999–9013, 2021.

[27] G. Liu, H. Yue, J. Wu, and J. Yang, "Efficient light field angular super-resolution with sub-aperture feature learning and macro-pixel upsampling," *IEEE Trans. Multimedia*, early access, Oct. 10, 2022, doi: 10.1109/TMM.2022.3211402.

[28] T. G. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. K. Nayar, and C. J. R. T. Intwala, "Spatio-angular resolution tradeoffs in integral photography," *Rendering Techn.*, vol. 2006, nos. 263–272, p. 21, 2006.

[29] 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.

[30] J. Shi, X. Jiang, and C. Guillemot, "Learning fused pixel and feature-based view reconstructions for light fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2555–2564.

[31] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong, and J. Yu, "Deep Coarse-to-Fine dense light field reconstruction with flexible sampling and geometry-aware fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1819–1836, Apr. 2022.

[32] N. Meng, K. Li, J. Liu, and E. Y. Lam, "Light field view synthesis via aperture disparity and warping confidence map," *IEEE Trans. Image Process.*, vol. 30, pp. 3908–3921, 2021.

[33] M. Guo, J. Jin, H. Liu, and J. Hou, "Learning dynamic interpolation for extremely sparse light fields with wide baselines," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2450–2459.

[34] 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. (ECCV)*, 2018, pp. 333–348.

[35] 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.

[36] G. Wu, Y. Liu, L. Fang, and T. Chai, "Revisiting light field rendering with deep anti-aliasing neural network," *IEEE Trans. Pattern Aanl. Mach. Intell.*, vol. 44, no. 9, pp. 5430–5444, Apr. 2021.

[37] M. Suhail, C. Esteves, L. Sigal, and A. Makadia, "Light field neural rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8269–8279.

[38] J. Yang, L. Wang, L. Ren, Y. Cao, and Y. Cao, "Light field angular super-resolution based on structure and scene information," *Appl. Intell.*, vol. 53, pp. 1–17, Feb. 2022.

[39] N. Meng, H. K.-H. So, X. Sun, and E. Y. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 43, no. 3, pp. 873–886, Mar. 2019.

[40] Z. Hu, Y. Y. Chung, W. Ouyang, X. Chen, and Z. Chen, "Light field reconstruction using hierarchical features fusion," *Exp. Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113394.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[42] Z. Hu, H. W. F. Yeung, X. Chen, Y. Y. Chung, and H. Li, "Efficient light field reconstruction via spatio-angular dense network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[43] Z. Cheng, Y. Liu, and Z. Xiong, "Spatial-angular versatile convolution for light field reconstruction," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 1131–1144, 2022.

[44] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 31–42.

[45] Y. Wang, L. Wang, G. Wu, J. Yang, W. An, J. Yu, and Y. Guo, "Disentangling light fields for super-resolution and disparity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 425–443, Jan. 2022.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[47] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understand.*, vol. 145, pp. 148–159, Apr. 2016.

SHUBO ZHOU received the B.E. and Ph.D. degrees from Beihang University (BUAA), China. He was a Postdoctoral Researcher with the Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), China, from 2017 to 2019. His currently an Assistant Professor with the Institute of Information Science and Technology, Donghua University, China. His research interests include computational photography and low-level vision.

XUE-QIN JIANG received the B.E. degree in computer science from the Nanjing Institute of Technology, Nanjing, China, and the M.S. and Ph.D. degrees in electronics engineering from Chonbuk National University, Jeonju, South Korea. He is currently a Professor with the Institute of Information Science and Technology, Donghua University, China. He has authored or coauthored more than 60 SCI articles. His research interests include wireless communications, quantum key distribution, signal processing, and machine vision. He is also serving as an Editor for IEEE COMMUNICATIONS LETTERS and IEEE ACCESS.

XIAOMING DING received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2012 and 2019, respectively. He is currently an Assistant Professor with Tianjin Normal University, Tianjin, China. His research interests include computational imaging, spectral light field imaging, and snapshot imaging spectrometer.

RONG HUANG received the B.S. degree in information engineering from the East China University of Science and Technology, Shanghai, China, in 2008, and the Ph.D. degree in advanced information technology from Kyushu University, Fukuoka, Japan, in 2013. He was with Kyushu University as an Academic Researcher, from October 2013 to March 2014. Since 2014, he has been with the Faculty of the College of Information Science and Technology, Donghua University, China, where he is currently an Associate Professor with the Institute of Information Science and Technology. His research interests include image processing, multimedia security, and machine learning.