Light Field Spatial Super-Resolution via Geometric Feature Interaction

Xin Chen, Yilei Chen, Ping An, Xinpeng Huang and Chao Yang

Light field (LF) enables high-dimensional image data representation since it can capture spatial and angular information of light rays simultaneously. The low spatial resolution caused by the limited imaging ability of the capturing equipment and the trade-off between spatial and angular resolution greatly affects the quality and application of LF images. In this letter, we propose an end-to-end LF super-resolution (SR) method via geometric feature interaction. Firstly, the low-resolution LF images are stacked in the horizontal and vertical epipolar plane image (EPI) directions and form 3D VI stacks. Then, these stacks are put into a dual-branch network, and we alternately perform 3D convolution on the viewpoint images (VIs) and EPIs by reshaping features for better feature extraction and interaction. The proposed method can fully explore the texture information and geometric consistency of the LF, and super-resolve all VIs at the same time. Experimental results on both real-world and synthetic LF datasets show that the proposed method has higher performance than other state-of-the-art methods.

Introduction: Light field (LF) imaging is one of the most widely used methods to capture the 3D appearance of scenes. Compared with traditional 2D images which capture the spatial information of light rays only, LF images can obtain both spatial coordinates and incident angles of light rays simultaneously [1]. Thus LFs enable lots of applications such as depth estimation [2], image refocusing [3] and 3D reconstruction [4]. However, due to the limited imaging ability of the capturing equipment and the trade-off between spatial and angular resolution, LF images have much lower spatial resolution than traditional 2D images. Therefore, in this letter, we focus on the further exploration of the LF spatial super-resolution (SR) technology.

Unlike the 2D image SR based on scene content prior, the pixel information needed for LF SR actually exists in each viewpoint image (VI). Conventional LF SR methods [5-7] explicitly warp the pixels of VIs using the disparity prior. These methods usually require complex optimization models such as Gaussian mixture model [5], variational model [6] and graph-based regularization model [7] to obtain better SR results. However, the occlusion and noise in LFs lead to the loss of valid pixel information, and disparity information heavily depends on the quality of the VI itself. Therefore, it is difficult for conventional LF SR methods to obtain high-quality SR results.

With the development of deep learning, convolutional neural networks (CNNs) are used for LF SR [8-15]. Yuan et al. [8] applied a single-image SR method to LF VIs, and then used another network to improve the quality of epipolar plane images (EPIs). But this method treated related VIs as separate individuals and ignored the connection between them. Zhang et al. [9] used VIs stacked in four directions to super-resolve the central VI, and there were different strategies for VIs with different angular locations. However, this method only used partial VIs for LF SR, and the remaining VIs were wasted. Recently, there are also methods of using all the information of VIs. Jin et al. [11] proposed an all-to-one method, which super-resolved each VI by taking all the remaining VIs as references. Wang et al. [12] used an ordinary convolution and a dilated convolution to extract spatial features and angular features, and then achieved LF SR through their interaction. Liang et al. [13] proposed a mixed-angular-resolution training strategy and a decouple-and-fuse module to achieve angular-flexible SR. However, these methods [10-14] lacked the use of geometric consistency of LF, which is characterized by EPI. On the basis of [9], Zhang et al. [15] used 3D convolution to extract the features of VIs and EPIs to accomplish LF SR. However, there are still some limitations on implicit learning of EPI information. In summary, learning-based methods can learn useful information for LF SR implicitly. However, they lack the interaction with geometric characteristics such as EPI. So the overall performance of LF SR still has room for improvement.

Inspired by the review, we propose an end-to-end LF SR method via geometric feature interaction. With the property that EPI information is contained in 3D VI stacks, we use a residual network with a dual-branch structure to learn the features of VIs, horizontal and vertical EPIs. Specifically, we use 3D convolution to alternately perform convolution on VIs and EPIs by reshaping features during convolution for better feature extraction and interaction. This enables the network to explore both the texture information and the geometric consistency of LFs. Experimental results on both real-world and synthetic LF datasets show that our method has higher objective performance than other state-of-the-art methods, and our results have more image detail and better geometric consistency. We also conduct ablation experiments to demonstrate the reasonableness and effectiveness of our method.

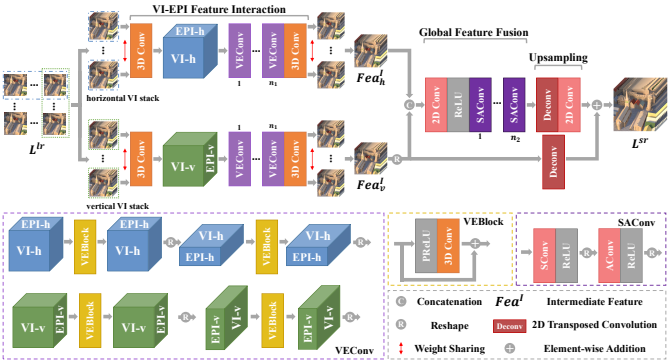

Fig. 1. Proposed network structure

Proposed method: With the property that EPI information is contained in 3D VI stacks, we propose an end-to-end method via geometric feature interaction by making full use of the information of VIs, horizontal and vertical EPIs. The network structure of our method is shown in Fig.1, which consists of three main modules: the VI-EPI feature interaction module, the global feature fusion module and the final upsampling module. Firstly, the low-resolution LF images are stacked in the horizontal and vertical EPI directions forming two sets of 3D LF data, and they are input into the VI-EPI feature interaction module with a dual-branch structure to obtain two intermediate features. Then the global feature fusion module is used for further fuse and optimization of these intermediate features. Finally, features pass through the upsampling module to output the SR results.

EPIs in VI Stacks. In practical applications, an LF is usually represented as a VI array and denoted by $L(s,t,x,y)$, where s and t are the angular dimensions, x and y are the spatial dimensions. The 2D slice obtained by sampling the LF with a fixed angle and spatial dimension is the EPI, in which the same visible object points in scenes from different VIs form a continuous straight line due to the disparity, as shown in Fig. 2. This line effectively reflects the geometric consistency within an LF. The EPI with fixed s and x is called the horizontal EPI and denoted by $L_{s^*x^*}(t,y)$, and the one with fixed t and y is called the vertical EPI and denoted by $L_{t^*y^*}(s,x)$. Similarly, the 3D VI stack with fixed s is called the horizontal VI stack and denoted by $S_{s^*}(t,x,y)$, and the one with fixed t is called the vertical VI stack and denoted by $S_{t^*}(s,x,y)$. As shown in Fig. 2, the horizontal (vertical) VI stack contains horizontal (vertical) EPI slices. Therefore, when the 3D VI stacks are alternately reshaped to VI slices and EPI slices during the 3D convolution process, the dual-branch network is able to learn all the information of VIs, horizontal and vertical EPIs, which helps achieve higher SR quality.

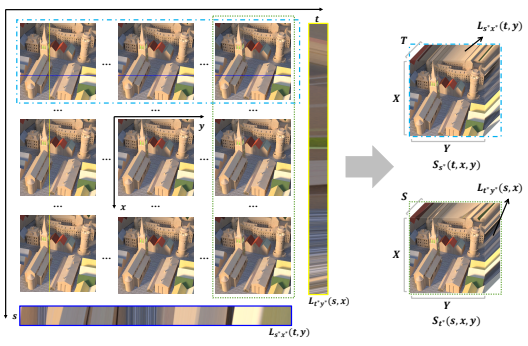

Fig. 2 *VI array, EPI (blue box for the horizontal EPI and yellow box for the vertical EPI) and 3D VI stacks stacked in EPI directions (light blue dotted box for the horizontal and green dotted box for the vertical VI stack), from which we can see that EPI information is contained in 3D VI stacks*

VI-EPI Feature Interaction. In order to make full use of the geometric consistency of the LF, the VI-EPI feature interaction module uses 3D convolution instead of 2D convolution for feature extraction. The 3D convolution on 3D VI stacks can perceive the information of VIs and EPIs simultaneously [15]. The size of the low-resolution LF is $S \times T \times X \times Y$, where $S \times T$ is the angular resolution, $X \times Y$ is the spatial resolution and S equals to T. When the low-resolution LF is stacked in the horizontal and vertical EPI directions, we can get two sets of 3D LF data, each of them contains S or T groups of 3D VI stacks. To extract features from both the horizontal and vertical EPI, the VI-EPI feature interaction module consists of horizontal and vertical branches. The input for the horizontal branch is S 3D VI stacks of size $N \times 1 \times A \times H \times W$, and for the vertical branch is T 3D VI stacks of size $N \times 1 \times B \times H \times W$, where N is the batch size. Furthermore, in each branch, the weights of the S (T) groups of 3D VI stacks are shared, and we use n_1 layers $VEConv$ to extract the interactive features of VIs and EPIs from the 3D VI stacks. The above convolution process is expressed by the formula as:

$$Fea_{s^*} = Conv_{h2}(VEConv_{hn_1}(\cdots VEConv_{h1}(Conv_{h1}(S^{lr}_{s^*})))) \quad (1)$$

$$Fea_{t^*} = Conv_{v2}(VEConv_{vn_1}(\cdots VEConv_{v1}(Conv_{v1}(S^{lr}_{t^*})))) \quad (2)$$

where h and v represent the horizontal and vertical EPI direction, $VEConv_i(\cdot)$ represents the i-th $VEConv$, $i \in 1, 2, \cdots n_1$, $Conv(\cdot)$ represents the 3D convolution, and $S^{lr}_{s^*/t^*}$ represents the low-resolution 3D VI stack stacked in the horizontal or vertical EPI direction.

As shown in Fig. 2, in 3D VI stacks of size $S \times X \times Y$and $T \times X \times Y$, the $X \times Y$ slice represents VI information, while the $X \times S$ and $T \times Y$ slices represent EPI information. For better extraction and interaction of features of EPIs and VIs, the features are alternately reshaped to VI slices and EPI slices during convolution in $VEConv$. Specifically, as shown in Fig. 1, the $VEConv$ is composed of $VConv$ and $EConv$, and both $VConv$ and $EConv$ are made up of resblock $VEBlock$. The $VConv$ represents the convolution on VIs, and the $EConv$ represents the convolution on EPIs. During this process, the feature size of the horizontal branch changes as follows:

$$S \times X \times Y \xrightarrow[reshape]{VConv} Y \times X \times S \xrightarrow[reshape]{EConv} \cdots \xrightarrow[reshape]{EConv} S \times X \times Y \quad (3)$$

Similarly, the feature size of the vertical branch changes as follows:

$$T \times X \times Y \xrightarrow[reshape]{VConv} X \times T \times Y \xrightarrow[reshape]{EConv} \cdots \xrightarrow[reshape]{EConv} T \times X \times Y \quad (4)$$

Finally, the output of the VI-EPI feature interaction module is two intermediate features from two EPI directions, and each of them consists of S Fea_{s^*} or T Fea_{t^*}:

$$Fea^I_h = \{Fea_{s^*} | s^* \in 1, 2, ..., S\} \quad (5)$$

$$Fea^I_v = \{Fea_{t^*} | t^* \in 1, 2, ..., T\} \quad (6)$$

Global Feature Fusion. To further integrate and optimize the learned information from 3D VI stacks of two EPI directions and improve the wholeness of the SR results, we use n_2 layers $SAConv$ to fuse and optimize the intermediate features obtained by the VI-EPI feature interaction module. Firstly, the intermediate features Fea^I_h and Fea^I_v are concatenated to a global feature Fea_g. Then, the angular dimension of Fea_g is moved to the batch dimension, and the feature channels of it are increased via a 2D convolutional layer. After this, the global feature size becomes $(N \cdot S \cdot T) \times C \times X \times Y$, where C represents feature channels. Then, the obtained global feature Fea_g is input into $SAConv$ [10], which consists of two 2D convolutional layers $SConv$ and $AConv$ sequentially. Similar to $VConv$ and $EConv$, $SConv$ represents the convolution on the spatial dimension of LFs, which is performed on features of size $(N \cdot S \cdot T) \times C \times X \times Y$, and $AConv$ represents the convolution on the angular dimension of LFs, which is performed on features of size $(N \cdot X \cdot Y) \times C \times S \times T$. The feature size is also reshaped between $SConv$ and $AConv$.

Upsampling. After fusion and optimization, the optimized global feature $\hat{Fea}_g$ is upsampled to the desired high-resolution spatial size via a 2D transposed convolutional layer. Then, the feature channels of it are reduced to 1 by a 2D convolutional layer to obtain the high-resolution residual information. After this, the feature size becomes $(N \cdot S \cdot T) \times C \times (\alpha \cdot X) \times (\alpha \cdot Y)$, where α is the upsampling factor. The final SR result is obtained by adding the high-resolution residual information and the intermediate feature of the vertical branch, and the reason for using the vertical branch to be the residual branch is explained in section Ablation Experiments:

$$L^{sr} = Conv_{res}(Deconv_{res}(\hat{Fea}_g)) + Deconv(Fea^{IR}_v) \quad (7)$$

where $Conv(\cdot)$ represents 2D convolution, $Deconv(\cdot)$ represents 2D transposed convolution, Fea^{IR}_v represents reshaping Fea^I_v to change the stacking arrangement order of the features, making it consistent with that of Fea^I_h.

Table 1: The number and categories of LFs in training and test datasets used in our method

Datasets	real-world datasets		synthetic datasets	
	STLFA	Kalantari et al.[16]	HCI new	HCI old
training	88	72	20	–
test	108	30	4	5

Experiments: The training and test datasets we used in this letter follow [11] and the details of them are shown in Table 1, including 2 real-world datasets Stanford Lytro LF Archive (referred to as STLFA) and Kalantari et al. [16], and 2 synthetic datasets HCI new and HCI old. The training and test LFs we used have the same angular resolution of 7×7. During the training stage, the low-resolution LF images for training are downsampled by bicubic and cropped to patches of size 64×64 and augmented by flipping horizontally or vertically, or rotating 90/180/270 degree randomly. All experiments are implemented in Pytorch with RTX 3090 GPU. The 2D and 3D convolutional layers in the network all have 64 filters with kernel size 3×3 or $3 \times 3 \times 3$ and zero-padding is applied to all convolutional layers. In network settings, we set $n_1 = 10$ and $n_2 = 6$. Besides, we use the Xaviers algorithm to initialize the weights of each convolutional layer, and the Adam optimizer to optimize the network. The batch size is set to 1, and the learning rate is initialized to 10^{-4} and halved every 2000 epochs. During the training, the network is supervised using the L1 loss function between the final SR result L^{sr} and the ground truth. We use peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate our method objectively. For subjective evaluation, the details of SR results and the EPI of the texture area are compared and displayed. Our source codes are available at https://github.com/wennychen11/LF-VEnet.

Comparative Experiments. We choose 7 comparative methods to evaluate our method, including 5 state-of-the-art LF SR methods resLF [9], LF-ATO [11], SA-Inter [12], LF-AFnet [13] and MEG-Net [15], a single image SR method VDSR [17] and a baseline method of single image SR bicubic. VDSR is retrained using all the LF VIs from the training datasets in Table 1. The quantitative results are illustrated in Table 2, from which we can conclude that our method outperforms other methods on both real-world and synthetic LF datasets for both $\times 2$ and $\times 4$ SR. The qualitative results are shown in Fig 3, from which we can see that the SR results of our method are closer to the ground truth, and have richer image detail, clearer image texture and edges than other methods. Furthermore, the EPIs of SR results produced by our method also have higher consistency. The straight lines formed by the disparity in EPIs have no disconnection or misconnection, and there is no obvious sawtooth.

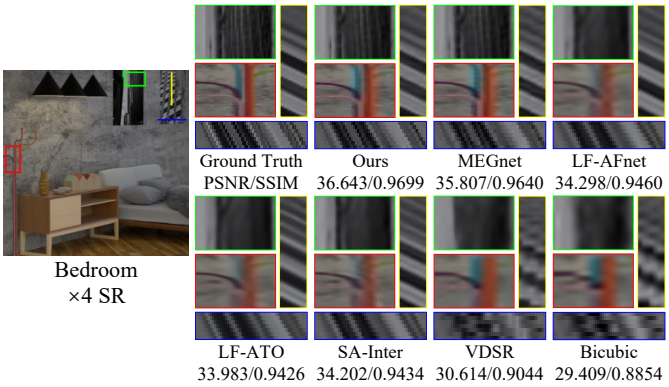

Fig. 3. Qualitative results of different methods for $\times 4$ SR

Ablation Experiments. We design some ablation experiments to demonstrate the reasonableness and effectiveness of our method, and the experimental results are shown in Table 3.

Table 2: Average PSNR/SSIM of results achieved by different methods for $\times 2$ and $\times 4$ SR. The data shown in bold is the best, and the sub-best data is shown in underline, the same as in Table 3

$\times 2$ SR	Bicubic	VDSR	resLF	LF-ATO	LF-AFnet	MEGnet	Ours
Kalantari et al.[16]	38.311/0.9839	40.686/0.9890	43.202/0.9937	44.025/0.9949	44.060/0.9951	44.105/0.9953	**44.666/0.9958**
HCI new	33.296/0.9525	35.016/0.9621	37.431/0.9784	38.519/0.9819	38.912/0.9839	38.196/0.9805	**39.493/0.9851**
STLFA General	36.464/0.9751	38.056/0.9806	40.558/0.9888	41.997/0.9916	41.838/0.9917	42.348/0.9927	**42.745/0.9931**
STLFA Occlusions	36.012/0.9744	38.151/0.9812	40.967/0.9888	41.916/0.9900	41.480/0.9877	41.294/0.9876	**42.048/0.9884**
HCI old	35.377/0.9655	37.208/0.9701	36.786/0.9609	39.499/0.9813	36.878/0.9526	40.514/0.9838	**41.245/0.9881**
$\times 4$ SR	Bicubic	VDSR	SA-Inter	LF-ATO	LF-AFnet	MEGnet	Ours
Kalantari et al.[16]	30.601/0.9215	32.089/0.9368	36.700/0.9719	36.911/0.9745	36.796/0.9746	37.634/0.9803	**37.747/0.9807**
HCI new	27.584/0.8632	28.848/0.8846	32.216/0.9303	32.276/0.9335	32.478/0.9369	32.862/0.9555	**33.755/0.9592**
STLFA General	29.564/0.9024	30.849/0.9187	34.839/0.9586	35.014/0.9613	34.744/0.9603	35.894/0.9712	**36.132/0.9724**
STLFA Occlusions	28.162/0.8765	29.323/0.8945	33.565/0.9415	33.878/0.9457	33.571/0.9434	34.011/0.9507	**34.253/0.9504**
HCI old	28.886/0.8677	30.312/0.8842	30.601/0.8745	32.751/0.9127	30.998/0.8793	34.973/0.9536	**35.438/0.9657**

Table 3: Average PSNR/SSIM of results achieved by several variants of ablation experiments for $\times 2$ SR

$\times 2$ SR	h-only	v-only	non-VE	$VVConv$	$EEConv$	$EVConv$	3D-only	ours
Kalantari et al. [16]	44.190/0.9954	44.187/0.9954	44.279/0.9954	44.583/**0.9958**	44.550/0.9957	44.636/**0.9958**	44.207/0.9954	**44.666/0.9958**
HCI new	38.930/0.9834	38.901/0.9832	39.017/0.9837	39.449/0.9850	39.371/0.9848	**39.498/0.9851**	38.822/0.9831	39.493/**0.9851**
STLFA General	42.193/0.9923	41.459/0.9880	42.257/0.9924	42.674/0.9930	42.654/0.9930	**42.771/0.9931**	42.215/0.9923	42.745/**0.9931**
STLFA Occlusions	41.448/0.9880	41.459/0.9880	41.616/0.9887	41.907/0.9872	41.936/0.9883	42.026/0.9875	41.494/0.9879	**42.048/0.9884**
HCI old	40.801/0.9864	40.883/0.9869	40.944/0.9871	41.202/0.9880	41.195/0.9878	41.237/0.9878	40.796/0.9867	**41.245/0.9881**

The dual-branch structure. We use horizontal-branch-only and the vertical-branch-only networks (referred to as h-only and v-only) to investigate the contribution of our dual-branch structure network. As shown in Table 3, the performance of these two variants is similar and much lower than our method. It proves that our dual-branch structure network can better explore all the information of VIs, horizontal and vertical EPIs. Meanwhile, v-only has a slight advantage than h-only, so we choose the vertical branch to be the residual branch in our network.

The $VEConv$ design. In our network, we design the structure of $VEConv$ to perform convolution on VIs and EPIs alternately. So we change the convolution order or type to show the effectiveness of this design. There are four variants called $VVConv$, $EEConv$, $EVConv$ and non-VE (delete all the $VEConv$ structures in our network). The experimental results show that $VEConv$ outperforms both $VVConv$ and $EEConv$, and has similar results to $EVConv$. Besides, non-VE has the worst performance in all variants. It can be demonstrated that the design of $VEConv$ can facilitate better extraction and interaction of features and improve the SR performance effectively.

The residual block. In $VEConv$, we use $VEBlock$ to achieve the convolution performed on 3D VI stacks. To prove its advantage, we use regular 3D convolutional layers to replace $VEBlock$ as a variant, and we call it 3D-only. The experimental results indicate that the residual structure $VEBlock$ can extract features of LFs more sufficiently.

Conclusion: In this letter, we propose an end-to-end LF SR method via geometric feature interaction. This method emphasizes the property that EPI information is contained in 3D VI stacks, so we use a residual network with a dual-branch structure to learn the information of VIs, horizontal and vertical EPIs from horizontal and vertical 3D VI stacks. Specifically, we use the $VEConv$ to perform convolution on VIs and EPIs alternately for better feature extraction and interaction. The experimental results show that our method outperforms other state-of-the-art methods in both objective metrics and subjective quality on both real-world and synthetic datasets. In the future, we will combine more characteristics of LFs to obtain higher quality SR results.

Acknowledgment: This work has been supported in part by the National Natural Science Foundation of China under Grants 62020106011, 62071287, 62001279 and 61901252, and by Science and Technology Commission of Shanghai Municipality under Grant 20DZ2290100.

Xin Chen, Yilei Chen, Ping An, Xinpeng Huang and Chao Yang (*School of Communication and Information Engineering, Shanghai Institute for Advanced Communication and Data Science Shanghai University, Shanghai, China*)

E-mail: anping@shu.edu.cn

References

1 Wu, G., Masia, B., Jarabo, A., Zhang, Y., Wang, L., Dai, Q., et al.: Light field image processing: an overview. *IEEE J. Sel. Top. Signal Process.* **11**(7), 926-54(2017)

2 Han, K., Xiang, W., Wang, E., Huang, T.: A novel occlusion-aware vote cost for light field depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 1-1(2021). http://doi.org/10.1109/TPAMI.2021.3105523

3 Jayaweera, S.S., Edussooriya, C.U., Wijenayake C., Agathoklis, P., Bruton, L.T.: Multi-volumetric refocusing of light fields. *IEEE Signal Process. Lett.* **28**, 31-35(2021)

4 Song, Z., Zhu, H., Wu, Q., Wang, X., Li, H., Wang, Q.: Accurate 3D reconstruction from circular light field using CNN-LSTM. In: IEEE Int. Conf. Multimed. Expo., pp. 1-6.(2020)

5 Mitra, K.,Veeraraghavan A.: Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior. In: Conf. Comput. Vis. Pattern Recognit. Workshops, pp. 22-28.(2012)

6 Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(3), 606-19(2014)

7 Ghassab, V.K., Bouguila, N.: Light field super-resolution using edge-preserved graph-based regularization. *IEEE Trans. Multimedia* **22**(6), 1447-57(2020)

8 Yuan, Y., Cao, Z., Su, L.: Light-field image superresolution using a combined deep CNN based on EPI. *IEEE Signal Process. Lett.* **25**(9), 1359-63(2018)

9 Zhang, S., Lin, Y., Sheng, H.: Residual networks for light field image super-resolution. In: Conf. Comput. Vis. Pattern Recognit., pp. 11038-47.(2019)

10 Yeung, H.W., Hou, J., Chen, X., Chen, J., Chen, Z., Chung, Y.Y.: Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Trans. Image Process.* **28**(5), 2319–30(2019)

11 Jin, J., Hou, J., Chen, J., Kwong, S.: Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In: Conf. Comput. Vis. Pattern Recognit., pp. 2257-66.(2020)

12 Wang, Y., Wang, L., Yang, J., An, W., Yu, J., Guo, Y.: Spatial-angular interaction for light field image super-resolution. In: Comput. Vis. ECCV, pp. 290-308.(2020)

13 Liang, Z., Wang, Y., Wang, L., Yang, J., Zhou, S.: Angular-flexible network for light field image super-resolution. *Electron. Lett.* **57**(24), 921-24(2021)

14 Ko, K., Koh, Y.J., Chang, S., Kim, C.-S.: Light field super-resolution via adaptive feature remixing. *IEEE Trans. Image Process.* **30**, 4114-28(2021)

15 Zhang, S., Chang, S., Lin, Y.: End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Trans. Image Process.* **30**, 5956-68(2021)

16 Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. *ACM Trans. Graph.* **35**(6), 193:1-193:10(2016)

17 Kim, J., Kwon, L.J., Mu, L.K.: Accurate image super-resolution using very deep convolutional networks. In: Conf. Comput. Vis. Pattern Recognit., pp. 1646–54.(2016)