Spatial-Angular Multi-Scale Mechanism for Light Field Spatial Super-Resolution

Chen Gao, Youfang Lin, Song Chang, Shuo Zhang*
Beijing Key Lab of Traffic Data Analysis and Mining
School of Computer and Information Technology, Beijing Jiaotong University
{gaochen, yflin, changsong, zhangshuo}@bjtu.edu.cn

Abstract

Light Field (LF) cameras are promising due to their ability to capture both spatial and angular information of scenes. However, the trade-off between spatial and angular resolution significantly limits the real-world applications. In this paper, we propose a spatial-angular multi-scale decoupling network to reconstruct high-resolution LF images. Considering the epipolar geometry, we propose a spatial-angular multi-scale processing approach to explore the correspondence of sub-pixel information with different disparity ranges between sub-aperture images in LFs. We extract sub-pixel information from various dimensions and fuse it to generate global high-frequency details. Finally, we combine upsampled low-frequency and high-frequency details to generate high resolution results. To further filter the correct interpolation information, we use the shear operation to change the disparity range of the LF images and fine-tune the results. Experimental results on synthetic and real-world datasets demonstrate that our method outperforms other state-of-the-art methods in visual and numerical evaluations, especially on datasets with small disparity ranges. Furthermore, our approach fully considers the epipolar geometry of the LF image, enabling us to recover information that better maintains the imaging consistency of the LF.

1. Introduction

Light Field (LF) cameras have immense potential applications in 3D reconstruction, virtual reality, and other fields due to their ability to capture both angular and spatial information of the scene [34]. In the past, LF images were captured using camera arrays that recorded Sub-Aperture Images (SAIs) with a large baseline. More recently, handheld plenoptic cameras [15, 17] have been developed by inserting a micro-lens array between the main lens and the imaging plane. These plenoptic cameras capture SAIs with a small baseline in a single shot, making them suitable for a wider range of applications, such as image refocusing [16]. However, the resolution of images generated by these cameras is lower than the large baseline LF cameras, limiting their effectiveness for many practical visual applications. To improve the performance of these applications, we need to enhance spatial resolution through Light Field Spatial Super-Resolution (LFSSR) technology.

Unlike a single image, LF images are represented by a sequence of multiple views to record scene information from different angles. Therefore, we can utilize the sub-pixel information between SAIs to recover High-Resolution (HR) details. In order to accurately find the position of sub-pixel in other SAIs, it is crucial to obtain disparity information for providing pixel-level offsets. Traditional methods have attempted to register sub-pixel information by explicitly warping other view images using prior disparity information [19, 29, 37]. However, existing methods for estimating disparity in LF images are susceptible to issues such as occlusions, noise, and textureless regions [6], which result in significant artifacts in the reconstructed LF images. Deep learning methods, particularly Convolutional Neural Networks (CNN), have recently been proposed to learn disparity information between SAIs for LFSSR implicitly [23, 32]. In order to explore the redundancy information between SAIs, CNNs learn the disparity information by processing different combinations of SAIs [5, 32] or processing information in both spatial and angular dimensions through convolution operations. Although LFSSR models have shown impressive performance in processing complex scenes, their effectiveness is limited to a specific range of scenes. Typically, these models perform better on specific disparity scenes, but their performance decreases significantly when dealing with large or small disparity ranges. In other words, the main challenge for current LFSSR models is to increase the search range of redundant information while maintaining their generalization ability.

After careful analysis of the phenomenon mentioned above, we have concluded that the decline in performance

*Corresponding author: zhangshuo@bjtu.edu.cn

can be attributed to the locality of the convolution operation. In LF images, sub-pixels are usually distributed widely. When there is a small disparity between scenes, it is often necessary to extract redundant information across multiple SAIs. However, when dealing with scenes that have a large disparity, the pixels move a significant distance in the spatial domain of adjacent SAIs. This opposite characteristic limits the potential of CNNs. To overcome this challenge, expanding the range of perceived information extracted by the convolution kernel is necessary, thus ensuring the complete extraction of sub-pixel information.

This paper proposes a Spatial-Angle Multi-Scale Spatial Super-Resolution (SAMSSR) network for LFSSR. In our method, we use the Multi-Dimension Interaction Block (MDIB) to process information from different organizational forms of LF images, particularly for extracting disparity information. In order to solve the problems existing in the current LFSSR method, we design the Multi-Scale Process Block (MSPB) to learn spatial-angular consistency information by processing multiple 2D Epipolar Plane Images (EPIs). Specifically, based on the multi-branch structure, we explicitly expand the perception range of disparity by setting different dilated convolutions to expand the receptive fields in the spatial and angular dimensions. Moreover, residual operations have been incorporated into our network. Furthermore, we propose a shear ensemble approach tailored with LF image Super-Resolution (SR) to enhance the SR performance with different disparity ranges.

Our proposed LFSSR method ranked well in the CVPR2023 NTIRE workshop challenge [24]. Extensive experiments over various challenging scenes show that the proposed SAMSSR achieves State-Of-The-Art (SOTA) results in terms of numerical and visual evaluations in LFSSR tasks. Moreover, the comparison of EPIs shows that the proposed method can preserve the corresponding relations in super-resolved view images. We also conducted additional experiments to validate the proposed ideas' effectiveness further. The contributions are:

1) We propose the MSPB, which expands the model's perception range to the LF images and solves the limitations of CNNs in processing sub-pixel information.

2) We proposed a shear ensemble approach tailored to LFSSR for performance enhancement.

3) Experiments show that our method expands the perceptual range of sub-pixel information and alleviates the performance degradation caused by the locality of convolution operation without increasing parameters.

2. Related Work

The learning-based LFSSR method can be divided into two categories: convolutional neural network-based models, and transformer-based models.

In recent years, deep CNNs have become widely used for Single-Image Super-Resolution (SISR) [3]. Various methods have been proposed, such as residual learning, recursive layers, or designing deeper convolutional models, resulting in significant performance improvements [7, 21, 39]. For LFSSR, CNNs have been used to learn the correspondence between LF SAIs and achieve interpolation of the correct pixels for SR tasks. For instance, LFCNN [33] utilizes SRCNN [2] to independently upsample the viewing angle and fine-tune the upsampled image in pairs. Yoon *et al*. [32, 33] designed spatial and angular SR networks to generate new SAIs by combining enhanced SAIs into different relative ones. Wang *et al*. [23] designed a bi-directional recurrent CNN for horizontal and vertical stack processing, which combines SAIs and utilizes stacked generalization techniques to produce a complete view image. To address the significant computational overhead of 4D convolution, Yeung *et al*. [31] proposed using Spatial-Angular Separable convolutions (SAS) to characterize LFs. Inspired by SAS, Wang *et al*. [26] proposed a new structure to extract spatial and angular features. Jin *et al*. [5] used the specific structure of the LF to propose an all-to-one structure and structural consistency regularization to guarantee the restoration of the characteristics of the image. Zhang *et al*. [36] used a multistream residual network by stacking SAIs along different angular directions as inputs and further improved the SR performance by performing 3D convolutions on SAI stacks of different angular directions [35]. These proposed LFSSR methods do not use both branches for processing but only extract sub-pixel information from the EPI structure. We think that disregarding the spatial and angle processing branches in LFSSR models results in a significant loss of information, which may negatively impact their performance. The spatial processing branch is capable of extracting the spatial features of SAIs as a whole, while the angular branch is adept at capturing angular information within fixed spatial pixels. This implicit learning of consistency between SAIs and their spatial features ensures the recovery of overall image style and light intensity. Thus, the spatial and angular processing branches are crucial for preserving important information and optimizing the performance of LFSSR models. Recently, Wang *et al*. [25] proposed a model for extracting different dimensional information of LF directly on micro-lens images.

The transformer model has also been applied to image-processing tasks, including LFSSR. Specifically, the transformer model is commonly utilized self-attention for modeling LF images and facilitating the transmission of cross-perspective information. This mechanism allows the model to effectively capture the inter-view dependencies in LF, which can help improve the quality of super-resolved images. Wang *et al*. [22] approach LFSSR as a task of re-

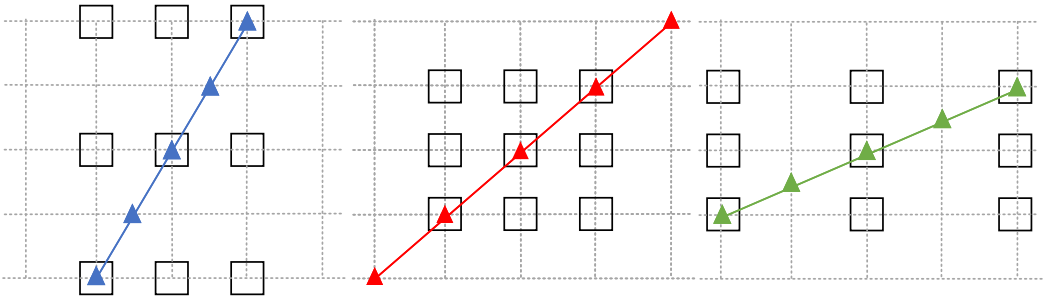

Figure 1. Display of three different disparity ranges of pixels on the horizontal EPIs. The intersection point of the dotted line grid represents the pixel on EPI, the three colors blue, red, and green represent the pixel points with three disparity ranges, and the square represents the convolution kernel. For pixels with different disparity ranges, different convolution kernels are needed to extract sub-pixel information.

constructing a sequence and propose a transformer model that preserves details by utilizing gradient maps of the LF to guide sequence learning. Afterward, Liang *et al*. [11] proposed a simple yet effective transformer-based method for LFSSR that captures both local and long-range dependencies within each SAI and incorporates complementary information among different SAIs. Recently, Liang *et al*. [12] proposed a transformer-based model named EPIT to learn non-local consistency by finding corresponding long-distance pixels. However, these methods do not consider different processing of sub-pixel information in different disparity ranges, which significantly affects the performance of LFSSR methods.

3. Motivation

ResLF [36] and MEG-Net [35] pointed out that EPI's rich sub-pixel information can help restore high-quality LFSSR images. That is because EPI has oriented line patterns with different slope values, which benefits estimating the disparity or reconstruction of LF in many challenging cases, such as occlusion and reflection areas.

Figure 1 illustrates three lines with different slopes corresponding to three disparity cases in the horizontal EPI. As mentioned, pixels with small disparity values move relatively slowly at adjacent angles, while those with large disparity values move relatively quickly in adjacent SAIs. We use blue, red, and green triangles to represent disparity values of 0.5, 1, and 2, respectively. When the angular coordinate moves, the three different pixel points will move 0.5, 1, and 2 pixels in the spatial dimension, respectively. As a result, we need to capture information from every other SAIs to obtain the corresponding pixel information for blue points. In the case of red disparity, we can find the necessary information from adjacent SAIs. However, green pixels have a large disparity value. Even a slight change in the angular coordinate can cause sub-pixel information to move considerably in the spatial dimension, spanning multiple spatial pixels. This non-local nature of pixel movement leads to the inability of the convolution kernel to fully extract the corresponding sub-pixel information when it performs feature extraction. To effectively capture the sub-pixel information of pixels with small disparity, extracting information from adjacent spatial coordinates across viewing angles is necessary. Conversely, for pixels with large disparity values, expanding the receptive field in the spatial dimension is essential. Thus, we use dilated convolution in the angular and spatial dimensions to enhance the SR model's performance.

In Fig. 1, we use a square to demonstrate the convolution kernel size used by most methods, which indicates the sub-pixel information range that can be perceived. The red pixels show that typical convolution can only perceive specific sub-pixel information ranges. Although the range of perception can be implicitly expanded through cascades and other methods, its effectiveness is limited. Our multi-scale strategy addresses this limitation by allowing convolution to perceive the sub-pixel information of blue and green points in different SAIs.

4. Methodology

LF is usually represented by a 4D tensor $\mathcal{L}(u, v, x, y) \in \mathbb{R}^{U \times V \times H \times W}$, where (u, v) represent angular coordinate, and (x, y) represent spatial coordinate [10]. Given the input LR image $\mathcal{L}^{\mathcal{LR}}$ of resolution (U, V, X, Y), our proposed model SAMSSR can extract the sub-pixel information from three representations of LF to restore the SR image $\mathcal{L}^{\mathcal{SR}}$ with resolution $(\alpha U, \alpha V, X, Y)$, where α represents the upsampling factor in spatial resolution.

4.1. Framework Overview

The overview of our SAMSSR is presented in Fig. 2 (a). Our network comprises three main stages: feature extrac-

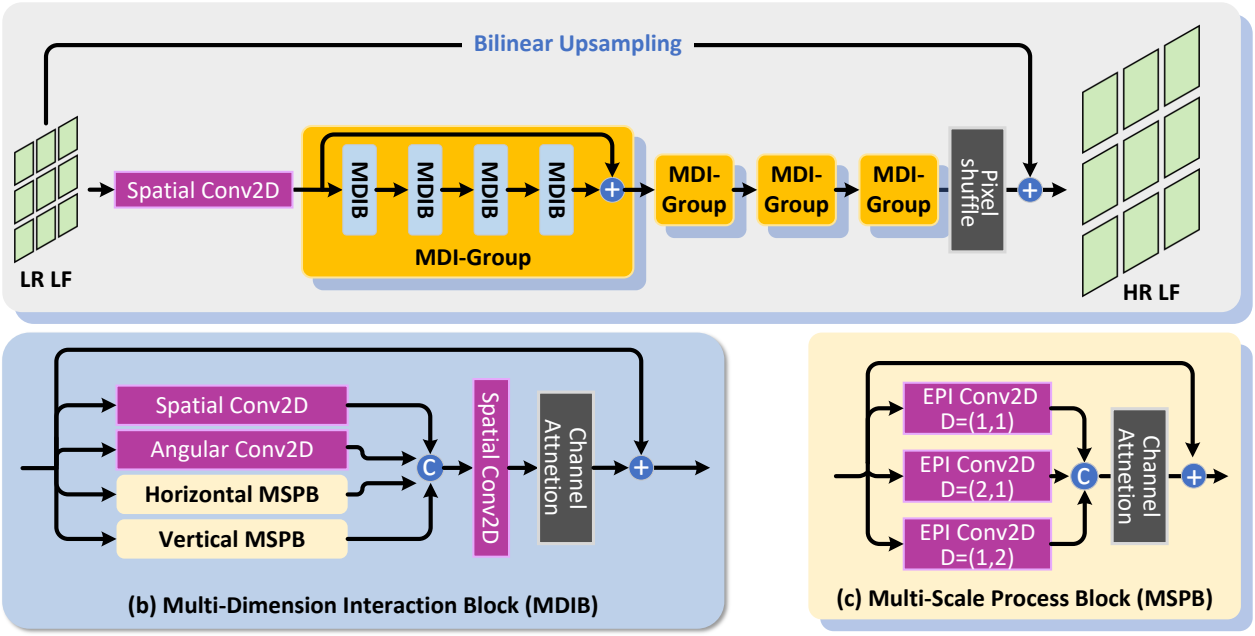

Figure 2. An overview of our SAMSSR network.

tion, spatial-angular decoupling, and feature upsampling. In the first stage, we extract features from the LR images. Next, the LR feature is passed through the cascaded Multi-Dimension Interaction Group (MDIG) module to extract sub-pixel information. The MDIG module decouples the spatial and angular information for more effective processing. After the high-frequency features are fully extracted, the output features are passed to the upsampling model to improve the image quality. In this paper, we convert images to YCbCr color space and only deal with Y channel images for better training.

4.2. Feature Extraction

Similar to most LFSSR methods, we apply convolutional operations on the spatial dimension of the input LR SAIs to obtain the LR high-dimensional feature $\mathcal{F}_{SAIs}^{\mathcal{LR}} \in \mathbb{R}^{U\times V\times H\times W\times C}$ as follows:

$$\mathcal{F}_{SAIs}^{\mathcal{LR}}\left[u,v,:,:,:\right]=\mathcal{W}_{\mathcal{S}}\otimes\mathcal{L}^{\mathcal{LR}}\left[u,v,:,:,:\right], \quad (1)$$

where $\mathcal{F}_{SAIs}^{\mathcal{LR}} \in \mathbb{R}^{U\times V\times H\times W\times C}$ denotes the LR high-dimensional feature, $\otimes$ denotes the convolution operation. In detail, the Spatial Conv2D contains a convolutional layer with a kernel size of 3×3, a stride of 1.

4.3. Multi-Dimension Feature Interact

In SAMSSR, we utilize the cascaded MDIG module to disentangle spatial and angular information as follows:

$$\mathcal{F}_{MDIG}^{m}=MDIG^{m}\left(\mathcal{F}_{MDIG}^{m-1}\right), m=1,...,M \quad (2)$$

where $MDIG^{m}$ denote the process of the m-th block MDIG and the $\mathcal{F}_{MDIG}^{m}$ represent the output feature of the m-th block MDIG. Note that, the $\mathcal{F}_{MDIG}^{0}$ is the LR high-dimensional feature $\mathcal{F}_{SAIs}^{\mathcal{LR}}$.

As illustrated in Fig. 2(b), like the overall framework structure of the model, MDIG is composed of the basic modules MDIB in series. The main idea of MDIB is to untangle spatial and angular information by processing different forms of LF images. Specifically, we parallel process SAI, MacPI, and EPI through four branches.

For SAI branches, we use the same operation as the initial feature extraction. The convolution operation on the SAIs is used to extract the overall spatial information from each SAI, in which only pixels located on the corresponding SAI will be processed simultaneously.

$$\mathcal{F}_{spa}^{\mathcal{LR}}\left[u,v,:,:,:\right]=\mathcal{W}_{spa}\otimes\mathcal{F}^{\mathcal{LR}}\left[u,v,:,:,:\right], \quad (3)$$

The dimension of feature $\mathcal{F}_{spa}^{\mathcal{LR}} \in \mathbb{R}^{U\times V\times H\times W\times \frac{C}{4}}$ is a quarter of the input feature.

Similarly, for the MacPI branch, we design the Angular Conv2D to obtain the angular information of the LF images. Unlike the SAI branch, where we apply convolution operations to the angular dimensions of LF images. The output feature dimension is also a quarter of the feature $\mathcal{F}^{\mathcal{LR}}$

$$\mathcal{F}_{ang}^{\mathcal{LR}}\left[:,:,x,y,:\right]=\mathcal{W}_{ang}\otimes\mathcal{F}^{\mathcal{LR}}\left[:,:,x,y,:\right], \quad (4)$$

where $\mathcal{F}_{ang}^{\mathcal{LR}} \in \mathbb{R}^{U\times V\times H\times W\times \frac{C}{4}}$ denotes the angular dimension features.

The SAI branch and MacPI branch in MDIB use convolution kernels of the same size that do not share weights. These kernels have a size of 3 and a stride size of 1. In the two remaining horizontal EPI branch and vertical EPI branch, we exploit the similarity of spatial pixels from different SAIs on the EPI structure using the designed MSPB. Details of the MSPB will be shown in Sec. 4.4.

Finally, the features obtained by the four branches are reshaped to form SAIs and concatenated. After that, we use a 1×1 convolution to reduce the channels of the concatenated feature. The channel attention module and the local residual learning operation obtain the final feature.

4.4. Spatial-Angular Mutil-Scale Process

In the horizontal EPI branch and the vertical EPI branch of the MDIB, we process the sub-pixel information by multi-scale operation in spatial and angular dimensions. The initial feature $\mathcal{F}^{\mathcal{LR}}$ is first reshaped into the horizontal EPI $\mathcal{F}_h^{\mathcal{LR}} \in \mathbb{R}^{VW \times U \times H \times C}$ and the vertical EPI $\mathcal{F}_v^{\mathcal{LR}} \in \mathbb{R}^{UH \times V \times W \times C}$. Then, feature $\mathcal{F}_H^{\mathcal{LR}}\left(\mathcal{F}_V^{\mathcal{LR}}\right)$ is fed into the MSPB module to interact with pixels with different disparity ranges.

$$\mathcal{F}_{epih}^{\mathcal{LR}} = MSPB\left(\mathcal{F}_H^{\mathcal{LR}}\right), \tag{5}$$

$$\mathcal{F}_{epiv}^{\mathcal{LR}} = MSPB\left(\mathcal{F}_V^{\mathcal{LR}}\right), \tag{6}$$

As shown in Fig. 2 (c), MSPB and MDIB are structurally similar and perform multi-branch parallel processing based on residual connection. However, MDIB focuses on feature extraction in different dimensions, and MSPB uses multi-scale operations to explore the differences between views to achieve accurate interpolation. For a given feature $\mathcal{F}_H^{\mathcal{LR}}\left(\mathcal{F}_V^{\mathcal{LR}}\right)$, we use dilated convolution kernels with different dilate rates. Each branch has the same convolution kernel size 3 but has different dilate rates of (1×1), (2×1), (1×2), respectively. As analyzed in Fig. 1, sub-pixel information in different disparity ranges is extracted only after the dilate operation is performed in a certain dimension of spatial or angular. Expanding in the angular dimension can ensure that the model extracts information on small disparity pixels while extending in the spatial dimension can ensure that the model extracts a larger range of disparity. Under the condition of ensuring the number of parameters, the accuracy degradation caused by the locality of the single convolution kernel size is alleviated. Note that the EPI structure can compare the relationship between the intuitive response spatial and angular. We let the model focus on processing the EPI structure so the output features of each branch have a higher feature dimension. Taking vertical EPI as an example, the process of three branches is as follows:

$$\mathcal{F}_{epiv}^{\mathcal{A}}[:, v, y, :] = \mathcal{W}_{epiv}^{\mathcal{A}} \otimes \mathcal{F}_V^{\mathcal{LR}}[:, v, y, :], \tag{7}$$

$$\mathcal{F}_{epiv}^{\mathcal{N}}[:, v, y, :] = \mathcal{W}_{epiv}^{\mathcal{N}} \otimes \mathcal{F}_V^{\mathcal{LR}}[:, v, y, :], \tag{8}$$

$$\mathcal{F}_{epiv}^{\mathcal{S}}[:, v, y, :] = \mathcal{W}_{epiv}^{\mathcal{S}} \otimes \mathcal{F}_V^{\mathcal{LR}}[:, v, y, :], \tag{9}$$

where $\mathcal{W}_{epih}^{\mathcal{A}}$, $\mathcal{W}_{epih}^{\mathcal{N}}$, $\mathcal{W}_{epih}^{\mathcal{S}}$ denote the convolution kernel parameter with dilated rates of (2×1), (1×1), (1×2). The channels for the three outputs are $\frac{3C}{8}$. Then, the features will be concatenated and fused by the channel attention module.

Although the structure of model [25] is similar to ours, their focus is different. While their model also extracts information from four dimensions, they emphasize extracting spatial and angular information, whereas we focus more on decoupling spatial and angular information. In addition, we propose a spatial-angular multi-scale process to implicitly expand the disparity range that the model can explore while reducing the number of parameters. Our method also avoids feature resolution reduction due to convolution operations during processing.

4.5. Feature Upsampling

After the model has extracted sufficient information in the four dimensions of the LR SAIs, we use the upsampling operation to restore the HR image as:

$$\mathcal{LF}^{SR} = UP\left(\mathcal{F}_{MDIG}^{M}\right), \tag{10}$$

where the UP denote the upsample model and the $\mathcal{LF}^{SR}$ means the final output HR LF image. Specifically, we use the pix-shuffle operation to recover the high-frequency information of the HR image. In order to ensure that the low-frequency information of the original image is preserved, we upsample the input image and obtain the final super-resolution result through residual learning. The final result is expressed as:

$$\mathcal{F}^{\mathcal{SR}} = \mathcal{F}_{up}\left(\mathcal{L}^{\mathcal{LR}}\right) + pixshuffle\left(\mathcal{F}_{MDIG}^{M}\right), \tag{11}$$

where $\mathcal{F}_{up}$ denote the bicubic operation.

4.6. Refine Mechanism

To ensure the applicability of the final model to a broader range of disparity, we incorporate the LF Shear Attention network [1] as a second-stage model to enhance the accuracy of the results. Firstly, the pre-trained SAMSSR model is applied to sheared LF images with varying disparity values, resulting in a set of SR outputs. These results are then sheared back using the $\times\alpha$ disparity values to restore the original disparity. Subsequently, we train the LF Shear Attention network [1] to identify relevant information from different sheared levels and fuse them to generate the final SR outcome.

5. Experiment

We use five public datasets EPFL [18], HCInew [4], HCIold [30], INRIA [9], STFgantry [20] for training and testing. All images have an angular resolution of 9×9. We calculate the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measurement (SSIM) for the Y channel of the image [28] as the evaluation metrics. We first demonstrate the performance of our approach in this challenge. We then conducted a comparative analysis between our SAMSSR model and several SOTA methods, includes three SISR methods (i.e., VDSR [7], EDSR [7], RCAN [38]) and ten LF image SR methods (resLF [36], LFSSR [31], LFATO [5], LF-InterNet [26], LF-DFnet [27], MEG-Net [35], LF-IINet [14], DPT [22], LFT [11], DistgSSR [25]). Finally, the ablation experiments were conducted to verify the effectiveness of the block in SAMSSR and the refine mechanism.

5.1. Implementation Details

Our network was trained using the L1 loss and optimized using the Adam method [8] with β_1=0.9, β_2=0.999. Our SAMSSR was implemented in PyTorch on a PC with one NVidia RTX A5000 GPU. The learning rate was initially set to 2×10^{-4} and decreased by 0.5 for every 20 epochs. The training was stopped after 65 epochs. Besides, we randomly augmented the datasets by flipping the images horizontally, vertically, or rotating 90 degrees.

For the first stage of the model SAMSSR training, we randomly crop successive 5×5 SAIs to improve the generalization ability of the model. For $4\times$ SR tasks, we crop the SAIs and use the bicubic downsampling to generate patches of size 32×32 as the LR input. Each MGI-group has four MDIB modules. Our SAMSSR consist of eight MGI-group modules.

Besides, in order to verify the validity of our proposed super-resolution model SAMSSR, we train the lightweight version, lablled as SAMSSR-l. Instead of randomly cropping 5×5 SAIs in 9×9 LFs, we crop the center 5×5 for training on the SAMSSR-l. Another difference is that SAMSSR-l has five MGI-group modules.

For the second stage of the refine model training, we add our data integration strategy to the above five datasets to generate training data. We perform model training and testing by cropping out the central 5×5 SAIs. We first downsample the image using bicubic interpolation and apply the shear operation based on the preset disparity values of $\{-1, -0.5, 0, 0.5, 1\}$. The trained SAMSSR model is used to upsample the sheared images, resulting in a set of SR outputs. These outputs are then sheared back using the disparity values of $\{4, 2, 0, -2, -4\}$ to generate the corresponding sheared image group as the input. Finally, we crop the SAIs into patches of size 128×128.

Table 1. Result of the LF image SR Challenge

Team Name	PSNR(avg)	#Params	Architec*	Rank
Group-1	30.664	20.34M	Hybrid	1
Group-2	30.6355	28.99M	CNN	2
Group-3	30.5619	10.52M	Transf	3
Group-4	30.3772	2.63M	Transf	4
SAMSSR+refine	30.3547	5.44M	CNN	5
Group-6	30.1286	8.83M	Transf	6
Group-7	30.1141	4.08M	Transf	7
Group-8	30.0559	3.35M	Transf	8
Group-9	29.8968	7.79M	CNN	9
Group-10	29.8492	14.82M	CNN	10
Group-11	29.8265	7.28M	CNN	11
Group-12	29.1163	-	-	12

5.2. Comparisons with Challenge methods

In this section, we present the performance of our method in the LFSSR challenge. We demonstrate the quantitative results in Tab. 1, which shows that our model (i.e., SAMSSR) achieves relatively good results while utilizing a small number of parameters. Furthermore, we provide the model architecture of the different methods in Tab. 1. Specifically, Transf denotes that the model adopts the Transformer as a basic component, CNN denotes that the model is developed based on convolutions only, and Hybrid denotes that the model contains sub-models developed using both CNNs and Transformers.

Among the models with the same architecture as ours, only the method in group two outperforms ours. However, the model parameters of this group are four times larger than ours, which would greatly affect computational efficiency. Moreover, our model performs better than some models based solely on the Transformer. This finding proves that the multi-branch structure of our model can effectively address the performance degradation caused by convolutional locality.

5.3. Comparisons with State-of-the-Art Methods

To ensure a fair comparison, we compare the performance of our model with the most recent state-of-the-art LF spatial SR methods, which have been trained on the same benchmark datasets.

Table 2 displays the quantitative results of SOTA methods on different upscaling factors. We use our lightweight SAMSSR-l model without the second refinement stage for a more fair comparison. Our SAMSSR-l model achieves the highest PSNR/SSIM values on nearly all datasets in both tasks, with a particularly significant improvement on the EPFI and INRIA datasets for $2\times$ SR. The main reason is that our method is able to capture in-depth features across the spatial, angular, and EPI domains. Additionally, the MSPB module extends its perceptual range on the

Table 2. Quantitative evaluations (PSNR / SSIM) of LFSSR results.The best results are in red, the second results are in blue.

Methods	#Params.	x2					x4				
	(×2/ × 4)	EPFL	HCInew	HCIold	INRIA	STFgantry	EPFL	HCInew	HCIold	INRIA	STFgantry
Bicubic	–	29.74/0.9376	31.89/0.9356	37.69/0.9785	31.33/0.9577	31.06/0.9498	25.26/0.8324	27.72/0.8517	32.58/0.9344	26.95/0.8867	26.09/0.8452
VDSR [7]	0.665M	32.50/0.9598	34.37/0.9561	40.61/0.9867	34.44/0.9741	35.54/0.9789	27.25/0.8777	29.31/0.8823	34.81/0.9515	29.19/0.9204	28.51/0.9009
EDSR [13]	38.62/38.89M	33.09/0.9629	34.83/0.9592	41.01/0.9874	34.99/0.9764	36.30/0.9818	27.83/0.8854	29.59/0.8869	35.18/0.9536	29.66/0.9257	28.70/0.9072
RCAN [38]	15.31/15.36M	33.16/0.9634	35.02/0.9603	41.13/0.9875	35.05/0.9769	36.67/0.9831	27.91/0.8863	29.69/0.8886	35.36/0.9548	29.81/0.9276	29.02/0.9131
resLF [36]	7.98/8.65M	33.62/0.9706	36.69/0.9739	43.42/0.9932	35.40/0.9804	38.35/0.9904	28.26/0.9035	30.72/0.9107	36.71/0.9682	30.34/0.9412	30.19/0.9372
LFSSR [31]	0.89/1.22M	33.67/0.9744	36.80/0.9749	43.81/0.9938	35.28/0.9832	37.94/0.9898	28.60/0.9118	30.93/0.9145	36.91/0.9696	30.59/0.9467	30.57/0.9426
LF-ATO [5]	1.22/1.36M	34.27/0.9757	37.24/0.9767	44.21/0.9942	36.17/0.9842	39.64/0.9929	28.51/0.9115	30.88/0.9135	37.00/0.9699	30.71/0.9484	30.61/0.9430
LF_InterNet [26]	5.04/5.48M	34.11/0.9760	37.17/0.9763	44.57/0.9946	35.83/0.9843	38.44/0.9909	28.81/0.9162	30.96/0.9161	37.15/0.9716	30.78/0.9491	30.37/0.9409
LF-DFnet [27]	3.94/3.99M	34.51/0.9755	37.42/0.9773	44.20/0.9941	36.42/0.9840	39.43/0.9926	28.77/0.9165	31.23/0.9196	37.32/0.9718	30.83/0.9503	31.15/0.9494
MEG-Net [35]	1.69/1.78M	34.31/0.9773	37.42/0.9777	44.10/0.9942	36.10/0.9849	38.77/0.9915	28.75/0.9160	31.10/0.9177	37.29/0.9716	30.67/0.9490	30.77/0.9453
LF-IINet [14]	4.84/4.89M	34.73/0.9773	37.77/0.9790	44.85/0.9948	36.57/0.9853	39.89/0.9936	29.04/0.9188	31.33/0.9208	37.62/0.9734	31.03/0.9515	31.26/0.9502
DPT [22]	3.73/3.78M	34.49/0.9758	37.35/0.9771	44.30/0.9943	36.41/0.9843	39.43/0.9926	28.94/0.9170	31.20/0.9188	37.41/0.9721	30.96/0.9503	31.15/0.9488
LFT [11]	1.11/1.16M	34.80/0.9781	37.84/0.9791	44.52/0.9945	36.59/0.9855	40.51/0.9941	29.26/0.9210	31.46/0.9218	37.63/0.9735	31.21/0.9524	31.86/0.9548
DistgSSR [25]	3.53/3.58M	34.81/0.9787	37.96/0.9796	44.94/0.9949	36.59/0.9859	40.40/0.9942	28.99/0.9195	31.38/0.9217	37.56/0.9732	30.99/0.9519	31.65/0.9535
SAMSSR-l (ours)	3.38/3.43M	35.18/0.9800	38.00/0.9800	45.06/0.9950	36.88/0.9860	40.60/0.9941	29.26/0.921	31.45/0.923	37.72/0.9740	31.23/0.9530	31.62/0.9534

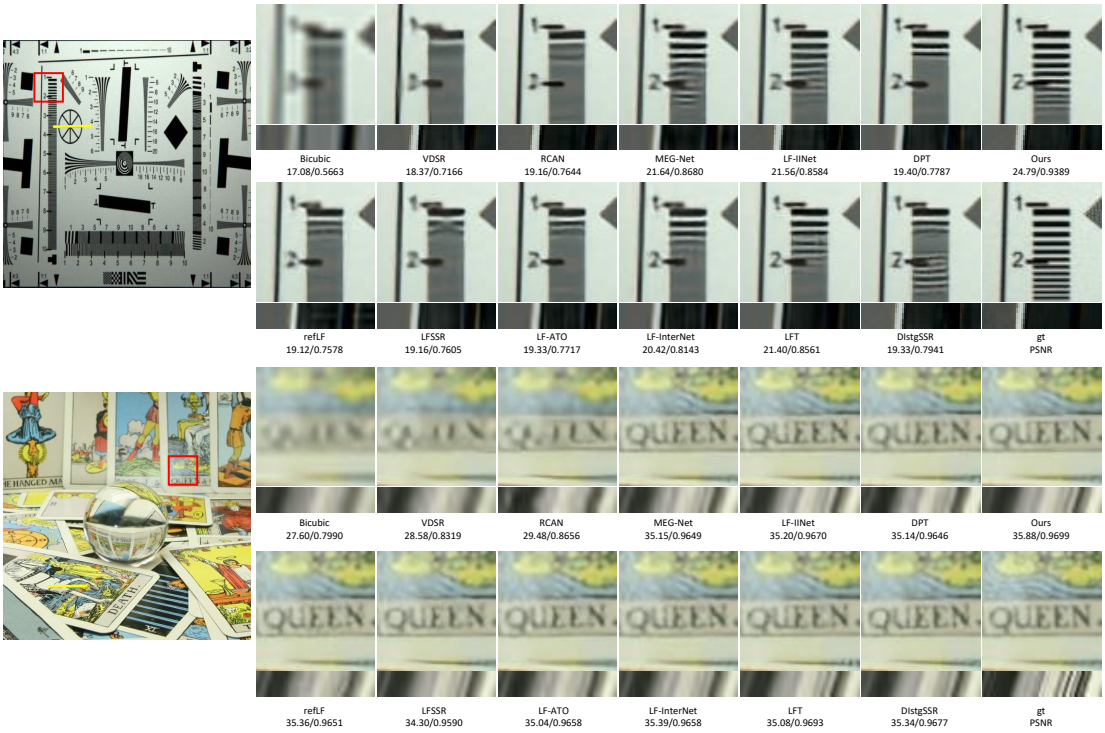

Figure 3. Visual comparison between the SOTA methods and our proposed method for an upscaling factor of ×4. Our results of the central view images and EPIs outperform the other SOTA methods with significant higher PSNR and SSIM. The background letters along the occlusion boundary are clearly recovered with sharp edges both in view and EPIs using the proposed method, while the others exhibit strong artifacts or ambiguous textures.

EPI dimension, enabling it to learn richer sub-pixel information. In contrast, previous methods do not thoroughly exploit all pixel correlations in 4D LF data, which results in limited performance gains. Notably, our model outperforms transformer-based models like LFT [22] and DPT [22] on most datasets, indicating that our MSPB partially overcomes the performance degradation caused by convolution operation locality. Especially on the small disparity datasets EPFL [18] and INRIA [9], our method outperforms the DistgSSR [25] methods by approximately 0.37dB and 0.29dB in terms of PSNR at a scaling factor of ×2.

Figure 3 offers a visual comparison between our method and other SOTA methods for an upscaling factor of ×4. We employ the SAMSSR and refine models to generate results. Our approach achieves superior perceptual quality in terms of both complex texture and detailed information when compared to other methods. Notably, our method recovers the exact shape of numbers with fine structures, while other

methods produce heavily blurred and broken characters in the "ISO_Chart_1" image from the EPFL dataset. This visual comparison aligns with the objective results, where our method outperforms the LFT [11] and DistgSSR [25] methods by approximately 3.39dB and 4.46dB, respectively, in terms of PSNR at a scaling factor of $\times 4$.

5.4. Ablation Investigation

In this subsection, we conduct experiments on 5×5 LF images for $4\times$ SR to investigate the effects of setting different dilation ratios in MSPB, attention fusion, and fine-tuning policies on the final results. We compared the variation models by averaging their performance across five datasets.

Table 3. Ablation experiments of dilation rate.

Spatial - Angular dilation in MSPB	Avg (PSNR/SSIM)
model $(\mathbf{1}-\mathbf{1})$	32.168/ 0.9444
model $(\mathbf{2}-\mathbf{2})$	32.254/ 0.9448
model $(\mathbf{2}-\mathbf{3})$	32.185/0.9443
model $(\mathbf{3}-\mathbf{2})$	32.175/0.9441
model $(\mathbf{3}-\mathbf{3})$	32.199/0.9446

The quantitative results of the different models with various dilation rates are presented in Tab. 3. For instance, in our MSPB module, the dilation rates of the three branch convolution kernels are $(\mathbf{2}\times 1)$, (1×1), and $(1\times \mathbf{2})$, which is labelled as model $(\mathbf{2}-\mathbf{2})$. We set up five models with varying dilation ratios of the spatial and angular domain in the EPI MSPB branches as model $(\mathbf{1}-\mathbf{1})$, $(\mathbf{2}-\mathbf{2})$, $(\mathbf{2}-\mathbf{3})$, $(\mathbf{3}-\mathbf{2})$, and $(\mathbf{3}-\mathbf{3})$. The best-performing model is achieved when the dilation rates of the spatial and angular dimensions on the EPI structure are set to 2. Compared with no dilation operation, i.e., model $(\mathbf{1}-\mathbf{1})$, our model has a wider sub-pixel information perception range, allowing for learning of more abundant redundant information and achieving accurate interpolation. Theoretically, when the dilation rate of the angular dimension is set to 3, the convolution kernel perceives the information among 7 views, which is needless for the 5×5 LF images. Similarly, when the dilation rate of the spatial dimension is set to 3, the convolution kernel perceives the information for scenes with disparity equal to 3, which is also needless for scenes with small disparities (for the $4\times$ SR task, the maximum disparity of the input image change is 1.75). Therefore, as the dilation rate increases, the effect of our model does not improve.

Besides, We demonstrate the effectiveness of the channel attention mechanism for multi-branch fusion by selectively removing them from our network. Specifically, "w/o-att1" meant removing the channel attention of the MDIB, while "w/o-att2" meant removing the channel attention of the MSPB. As shown in Tab. 4, the performance of the models declined when we remove the channel attention mechanism from both models. Notably, when we remove the attention mechanism in the MSPB module, the model performs poorly on the datasets, indicating that the branch information fusion of the MSPB module is insufficient, leading to the model being unable to adapt well to the extraction of sub-pixel information in multiple disparity ranges. Furthermore, when we remove the attention mechanism in the MDIB module, the model's performance declined on all datasets, highlighting the importance of the fusion of different forms of feature information from LF images.

Table 4. Ablation experiments of attention and refine strategy.

Variants	#Params	Avg (PSNR/SSIM)
w/o-att1	3.33M	32.184/0.9446
w/o-att2	3.39M	32.203/0.9444
SAMSSR-l	3.58M	32.254/ 0.9448
SAMSSR	5.43M	32.322/0.9451
SAMSSR+refine	5.44M	32.94085/0.9481

By comparing the performance of SAMSSR-l and SAMSSR, we observe that when we increase the basic module of the model (i.e., MDIB), the model generalizes better on the training data and improves the performance. Additionally, when we apply the refine strategy to the results of SAMSSR processing, the model performs significantly better with a small number of parameters. By using shear-based data integration operations, we vary the disparity range of the scene and train models to better extract interpolating information from SR images of LF with different disparity ranges in the same scene.

6. Conclusion

Our paper introduces the SAMSSR network as a solution to enhance the spatial resolution of light field images by leveraging the information extracted from spatial, angular, and epipolar feature extractors. The MSPB module, designed explicitly for EPI structures, extracts sub-pixel information. Our experimental results demonstrate that our method outperforms state-of-the-art methods in terms of PSNR and SSIM and delivers superior visual effects. Additionally, our network effectively preserves the epipolar property of images and can be applied to various types of light field images with different angular resolutions. Meanwhile, our approach ranked fifth in the Light Field Image Super-Resolution challenge.

Acknowlegement

This work was supported by the National Natural Science Foundation of China under Grant 61902018.

References

[1] Yangling Chen, Shuo Zhang, Song Chang, and Youfang Lin. Light field reconstruction using efficient pseudo 4d epipolar-aware structure. *IEEE Transactions on Computational Imaging*, 2022. 5

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 184–199, Cham, 2014. Springer International Publishing. 2

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 2

[4] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 19–34. Springer, 2016. 6

[5] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2260–2269, 2020. 1, 2, 6, 7

[6] Ole Johannsen, Katrin Honauer, Bastian Goldluecke, Anna Alperovich, Federica Battisti, Yunsu Bok, Michele Brizzi, Marco Carli, Gyeongmin Choe, Maximilian Diebold, et al. A taxonomy and evaluation of dense light field depth estimation algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 82–99, 2017. 1

[7] Jiwon Kim, Lee Jung Kwon, and Lee Kyoung Mu. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 2, 6, 7

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[9] Mikael Le Pendu, Xiaoran Jiang, and Christine Guillemot. Light field inpainting propagation via low rank matrix completion. *IEEE Transactions on Image Processing*, 27(4):1981–1993, 2018. 6, 7

[10] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996. 3

[11] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. Light field image super-resolution with transformers. *IEEE Signal Processing Letters*, 29:563–567, 2022. 3, 6, 7, 8

[12] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, Shilin Zhou, and Yulan Guo. Learning non-local spatial-angular correlation for light field image super-resolution, 2023. 3

[13] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 136–144, 2017. 7

[14] Gaosheng Liu, Huanjing Yue, Jiamin Wu, and Jingyu Yang. Intra-inter view interaction network for light field image super-resolution. *IEEE Transactions on Multimedia*, 2021. 6, 7

[15] Lytro. Lytro redefines photography with light field cameras. http://www.lytro.com, 2016. Accessed: Oct. 22, 2018. 1

[16] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, Pat Hanrahan, et al. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005. 1

[17] Raytrix. Raytrix: Light filed technology. http://www.raytrix.de, 2016. Aug. 22, 2020. 1

[18] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *8th International Conference on Quality of Multimedia Experience*, 2016. 6, 7

[19] Mattia Rossi and Pascal Frossard. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Transactions on Image Processing*, 27(9):4207–4218, 2018. 1

[20] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. The stanford lytro light field archive. http://lightfields.stanford.edu/LF2016.html, 2016. Accessed: Aug. 22, 2020. 6

[21] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5, 2017. 2

[22] Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Detail-preserving transformer for light field image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2522–2530, 2022. 2, 6, 7

[23] Yunlong Wang, Fei Liu, Kunbo Zhang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Transactions on Image Processing*, 27(9):4274–4286, 2018. 1, 2

[24] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 2

[25] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. 2, 5, 6, 7, 8

[26] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *Computer Vision – ECCV 2020*, pages 290–308, Cham, 2020. Springer International Publishing. 2, 6, 7

[27] Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo. Light field image super-resolution using deformable convolution. *IEEE Transactions on Image Processing*, 30:1057–1071, 2021. 6, 7

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[29] Sven Wanner and Bastian Goldluecke. Spatial and angular variational super-resolution of 4d light fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 608–621. Springer, 2012. 1

[30] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modeling & Visualization*, volume 13, pages 225–226. Citeseer, 2013. 6

[31] Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018. 2, 6, 7

[32] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24(6):848–852, 2017. 1, 2

[33] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE international conference on computer vision workshops (ICCVW)*, pages 24–32, 2015. 2

[34] Jingyi Yu, Xu Hong, Jason Yang, and Yi Ma. Dgene: The light of science, the light of future. http://www.plex-vr.com/product/model/, 2018. Accessed: Aug. 22, 2020. 1

[35] Shuo Zhang, Song Chang, and Youfang Lin. End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Transactions on Image Processing*, 30:5956–5968, 2021. 2, 3, 6, 7

[36] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11046–11055, 2019. 2, 3, 6, 7

[37] Shuo Zhang, Hao Sheng, Da Yang, Jun Zhang, and Zhang Xiong. Micro-lens-based matching for scene recovery in lenslet cameras. *IEEE Transactions on Image Processing*, 27(3):1060–1075, 2017. 1

[38] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 6, 7

[39] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2