Optica Applicata, Vol. LVI, No. 1, 2026
DOI: 10.37190/oa/216305

Light field image super-resolution network based on multi-dimensional disparity mining

LI HUANG, JINYAO ZHANG, JIANI CHEN, YINGCHUN WU*

School of Electronic Information Engineering, Taiyuan University of Science and Technology, No. 66 Waliu Road, Taiyuan 030024, China

*Corresponding author: yingchunwu3030@foxmail.com

Based on the two-plane representation model of the light field, light field cameras can synchronously record positional and directional information of the rays by sacrificing spatial resolution. To improve the spatial resolution of light field images, this paper presents a light field super-resolution network built upon the mechanism of multi-dimensional disparities mining. In the shallow layers of the network, a multi-scale self-attention module is designed, which works in conjunction with a residual atrous spatial pyramid pooling module to extract features from sub-aperture images. For the deep feature extraction, a multi-dimensional feature separator-mixer is developed to mine disparities from different forms of light field subspace, and an iterative structure is employed to extract deep fused disparity among different sub-aperture images. At the end of the network, a global-local feature refine module is added to further improve the quality of the output images. By adopting the mechanism of disparity iterative separation-fusion and re-enhancement, the proposed network achieves high-quality super-resolution reconstruction of sub-aperture images with a relatively compact architecture. Experimental results verify the superiority of the proposed network. In particular, with only 3.83 M parameters, the network achieves 4 times super-resolved images with an average PSNR of 32.27 dB in 5 public datasets.

Keywords: light field, spatial super-resolution, super-resolution network, disparity mining.

1. Introduction

Based on the light field (LF) two-plane parameterization model, a plenoptic camera simultaneously captures spatial and angular light ray information. Decoding 4D LF from a single exposure enables concurrent acquisition of multi-focus [1, 2], multi-view [3], all-in-focus [4, 5] images and depth maps [6-8]. However, recording angular information inevitably sacrifices spatial resolution, making 4D LF super-resolution (SR) reconstruction research highly significant. LF spatial SR aims to improve the spatial resolution of all angular images in the sub-aperture image (SAI) array, requiring exploitation of both individual SAI's internal structure and texture (spatial cues) and cross-SAI complementary information (angular cues). Due to the high coupling of spatial and angular cues, recent works have designed various advanced deep-learning

networks to mine these cues from multiple perspectives, steadily enhancing LF SR performance.

With an emphasis on jointly exploiting spatial and angular cues, YEUNG *et al*. proposed LFSSR [9], which represents pixel relationships with 4D convolutions and explores spatial-angular correlations by space-angle separable convolutions. LIANG *et al*. introduced LF-AFnet [10], it designs tailored feature extractors and a decoupling and fusion module for LF SR, and adopts a mixed-angular-resolution training strategy to endow the network with angular flexibility. LU *et al*. developed DPT [11], treating LF spatial SR as a sequence-to-sequence task guided by gradient maps. WANG *et al*. presented LF-Internet [12], which leverages a spatial-angular information interaction mechanism to make full use of the 4D LF structure. Building on this, they further proposed DistgSSR [13] that jointly learns spatial, angular and epipolar-plane image (EPI) features on the macro-pixel image and fuses them for comprehensive LF mining. YUE *et al*. devised LF-IInet [14], equipped with intra-view and inter-view mutual-update modules and 3D convolutions based multi-view context block, it models cross sub-aperture images (SAIs) dependencies while preserving parallax structure. They subsequently introduced LF-DGnet [15], a self-supervised framework that estimates disparity maps and uses them to modulate LF features, fully exploiting the correlations of the SAIs.

4D LF matrices contain not only rich spatial and angular information, but also disparity information. Disparity refers to the imaging differences of the same object across viewpoints, reflecting the positional discrepancies among SAIs. The sub-pixel information is crucial for LF spatial SR task. Based on this, this paper proposes a multi-dimensional disparity mining network for LF image SR. In the shallow layers of the network, residual atrous spatial pyramid pooling (ResASPP) block and multi-scale self-attention module are employed to extract SAI's information. For disparity mining, a multi-dimensional feature separator-mixer is designed under the iterative structure, by cyclically separating and mixing features, and it better explores disparity information across SAIs. A feature re-optimization stage is appended at the network's end to further boost performance. The main contributions of this paper are as follows:

(1) For sub-aperture feature extraction, this paper introduces a multi-scale self-attention module (MSM). By performing upsampling and downsampling on the features, MSM controls the receptive field and acquires self-attention weights at multiple scales, boosting feature utilization.

(2) To mine disparity cues, this paper proposes a multi-dimensional feature separator-mixer (MFSM). It extracts disparity features from global, local, horizontal and vertical perspectives and then fuses them to enrich disparity structures. Embedded in the iterative topology, MFSM is repeatedly stacked with equal-stride residual connections, enabling cyclic separation and fusing while preventing feature degradation.

(3) At the network tail, a global-local feature refine (GLFR) module is appended for feature re-optimization. Operating on the high-resolution feature map, GLFR extracts global and local information separately and fuses them, jointly updating coarse-grained and fine-grained details.

2. Network architecture

The imaging model of a LF camera relies on the two-plane parameterization of the 4D LF [16]. In a microlens-based LF camera, the two planes are defined as the main lens plane ($u\,v$) and the sensor plane ($s\,t$). In a camera-array LF system, the array itself forms the $u\,v$ plane, while the common focal plane of the cameras acts as the $s\,t$ plane. Any ray in space intersects these two planes at coordinates (u, v) and (s, t), respectively, and is thus denoted as $L(u, v, s, t)$, as illustrated in Fig. 1(a). The 4D LF is therefore represented by the tensor $L \in R^{U\times V\times S\times T}$, where $U\times V$ is the angular resolution and $S\times T$ is the spatial resolution. Two canonical visualizations of this 4D data are the raw microlens image array and the SAI array. Figure 1(b) shows the microlens image array. When magnified, it reveals the hexagonal grid imposed by the microlens sheet. After redundancy removal, each hexagon is termed a macro-pixel. Extracting the gray value at the same position inside every macro-pixel and rearranging these values yields the SAI array, displayed in Fig. 1(c).

The overall network architecture is shown in Fig. 2(a) and consists of four stages, namely sub-aperture feature extraction, disparity feature mining, upsampling, and feature re-optimization. Sub-aperture feature extraction is performed by ResAspp block followed by MSM. For multi-dimensional disparity mining, MFSM is adopted and ar-

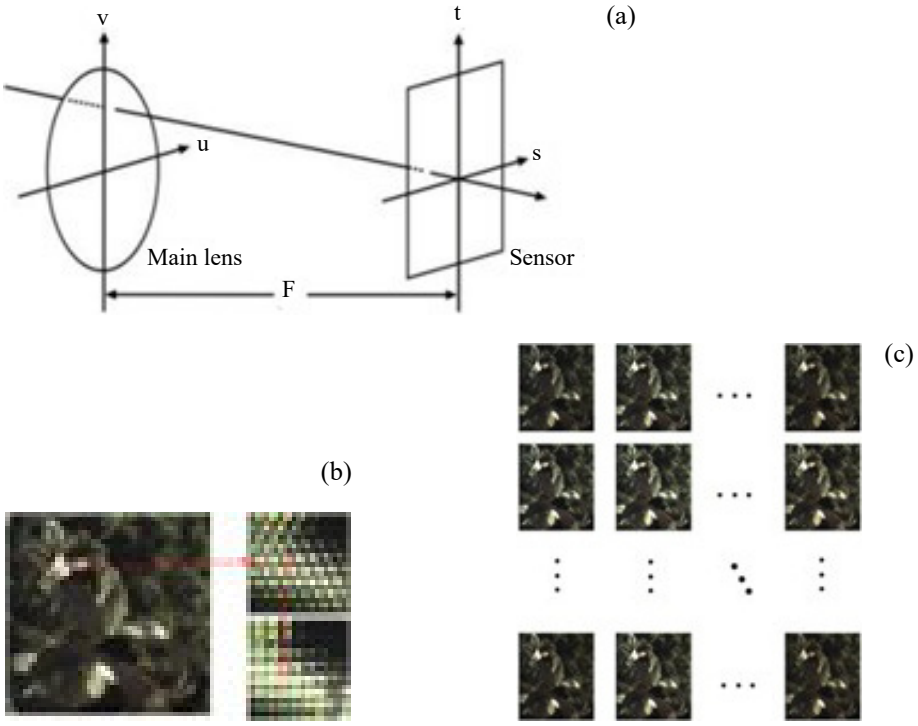

Fig. 1. 4D LF representation and visualization. (a) Two-plane parametrization model, (b) microlens image array, and (c) the SAI array.

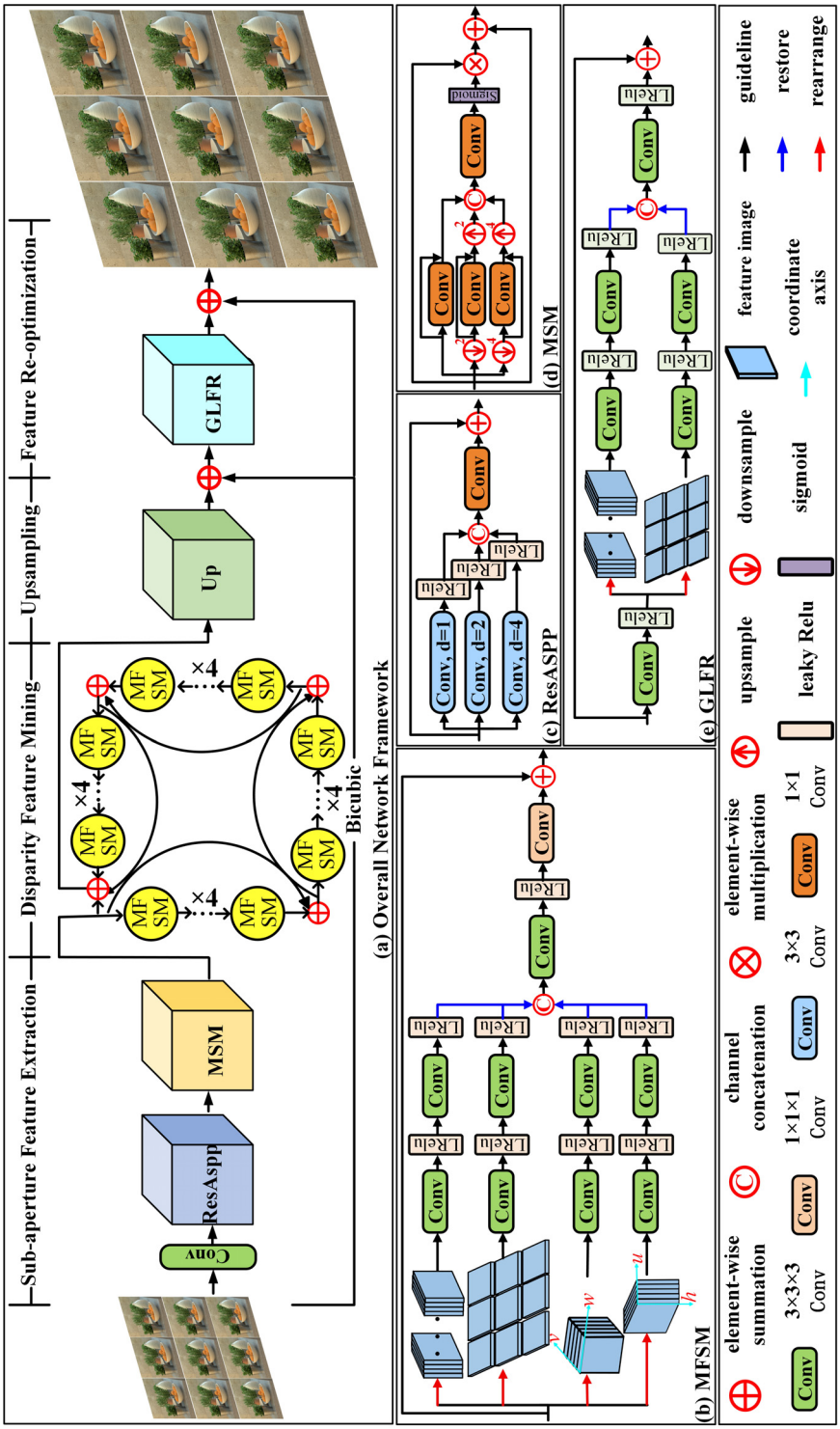

Fig. 2. An overview of the network. (a) Overall network framework, (b) MFSM, (c) ResASPP, (d) MSM, and (e) GLFR.

ranged in an iterative structure, it cyclically separates and fuses features to fully exploit the LF parallax structure. The upsampling stage decodes the low-resolution feature into a high-resolution SAI array. Finally, a re-optimization stage with the GLFR module operates on the high-resolution image, fusing global and local cues to further enhance details.

2.1. Sub-aperture feature extraction

In the shallow feature extraction stage of the proposed LF SR network, the SAI array $I^{LR} \in R^{U \times V \times S \times T}$ is first fed into a 3×3 convolution layer to produce elementary features. We then employ ResAspp and MSM to fully exploit the SAI's information, with both modules adopting a parallel strategy for aggregating multi-scale information. ResAspp captures multi-scale context by varying the dilation rates. MSM obtains multi-scale information through feature upsampling and downsampling.

The detailed operations of ResAspp and MSM are as follows. ResAspp first applies three parallel 3×3 atrous convolutions with rates 1, 2, 4, each followed by LeakyReLU, and their outputs are rearranged along the channel dimension. Then the number of channels is compressed by 1×1 convolution, and merged with the input via a residual link. MSM first splits the input feature into three paths. Path-1 keeps the original scale, path-2 operates on a 2× down-sampled version, and path-3 on a 4× down-sampled version. Each path is processed by a 1×1 depth-wise convolution with a residual shortcut. The outputs of path-2 and path-3 are then up-sampled back to the original resolution, concatenated with path-1 along the channel dimension, squeezed to a single channel by a 1×1 convolution, and normalized with a sigmoid to produce the self-attention weight. The final output is obtained by element-wise multiplication with the input feature to obtain an attention map, which is then added back to the input via a residual connection, easing gradient vanishing and counters network degradation. The detailed pipeline of MSM is illustrated in Fig. 2(d), the entire procedure can be expressed as follows:

$$F_{\mathrm{E}} = H_{\mathrm{MSM}}(H_{\mathrm{ResAspp}}(H_{3\times3}(I^{LR}))) \tag{1}$$

where $H_{\mathrm{ResAspp}}(\cdot)$ denotes the ResAspp operation, $H_{\mathrm{MSM}}(\cdot)$ denotes the MSM operation, $H_{3\times3}(\cdot)$ denotes the 3×3 convolution operation and F_{E} represents the extracted sub-aperture features.

2.2. Disparity feature mining

Multi-dimensional disparity mining aims to fully exploit parallax cues among different SAIs. Then, due to the fact that SAIs are regularly arranged along the angular dimension with distinct information, aiming to fully preserve the parallax structure of the light field, this paper proposes the MSFM to extract and fuse parallax features from different forms of light field subspace, yielding richer disparity representations. This design enables the generation of richer disparity representations, where the term "multi -dimensional" specifically refers to disparity features across these four distinct dimen-

sions. Specifically, Global disparity is obtained by stacking all SAIs and then applying convolution operations to fuse information along the channel dimension. Local disparity is acquired by concatenating adjacent SAIs and then using convolution operations to aggregate information along the spatial dimension. Global disparity focuses on the overall disparity correlation of the scene, while local disparity emphasizes the disparity correlation between adjacent SAI. Horizontal disparity is extracted by slicing the EPI volume with fixed U and S coordinates. Vertical disparity is obtained by slicing the EPI volume with fixed V and T coordinates. The global and local disparities in the proposed "multi-dimensional" feature correspond to the spatial information of light field, whereas the horizontal and vertical disparities correspond to the angular information. The extraction of features across four dimensions effectively decouples the spatial and angular information of light field, thus yielding favorable super-resolution performance. Stacking MSFM in an iterative structure with equal-stride residual connections enables cyclic separation-fusion of disparity features while preventing network degradation.

The detailed structure of MSFM is shown in Fig. 2(b). During feature separation, it contains four branches. The first branch takes the SAI array rearranged along the channel dimension as input, applies two $3\times3\times3$ convolutions and the LeakyReLU activation function to capture global disparity. The second branch operates directly on the original SAI array, and acquires local disparity via the same convolution and activation. The third branch feeds the horizontal EPI through the same operations to extract horizontal parallax. The fourth branch takes vertical EPI information as input, and acquires vertical disparity via the same operations. During feature fusion, the four-disparity information are concatenated along the channel dimension, forwarded through Conv-LeakyReLU-Conv, and finally added back to the input via a residual shortcut, yielding the module's output. This residual path preserves fine textures while combating degradation, and the diversified inputs allow MSFM to harvest parallax cues from multiple perspectives. The entire procedure can be written as:

$$\begin{cases} F_1 = H_{\mathrm{MSFM}\times4}(F_{\mathrm{E}}) + F_{\mathrm{E}} \\ F_2 = H_{\mathrm{MSFM}\times4}(F_1) + F_1 \\ F_3 = H_{\mathrm{MSFM}\times4}(F_2) + F_2 \\ F_{\mathrm{D}} = H_{\mathrm{MSFM}\times4}(F_3) + F_3 + F_{\mathrm{E}} \end{cases} \tag{2}$$

where $H_{\mathrm{MSFM}}(\cdot)$ denotes the MSFM operation, F_1, F_2, F_3 represents the output features of the first, second and third residual connections in the iterative structure, respectively. F_{D} represents the output feature of the disparity feature mining.

2.3. Upsampling

The upsampling stage maps low-resolution features to the high-resolution LF image. A 1×1 convolution first expands the channel dimension and the pixel shuffle transfor-

mation for mapping channel dimension to spatial dimension, yielding the final high-resolution LF image. The whole procedure can be expressed as follows:

$$F_{\mathrm{HR}} = H_{\mathrm{UP}}(F_{\mathrm{D}}) \tag{3}$$

where $H_{\mathrm{UP}}(\cdot)$ denotes the upsampling operation, F_{HR} represents the output of the high-resolution LF image.

2.4. Feature re-optimization

To further update coarse-and fine-grained details, this paper appends a feature re-optimization stage that operates directly on the high-resolution space, where a GLFR module is designed to separately extract and fuse global and local information from high -resolution images. The detailed structure of GLFR is shown in Fig. 2(e). The high-resolution SAI array from the upsampling module first undergoes a $3\times3\times3$ convolution and LeakyRelu activation, which initially explores the correlation between spatial and angular information. The feature is then split into two paths. The upper branch takes the channel-reordered SAI as input and applies two $3\times3\times3$ convolutions with LeakyReLU to capture global disparity, while the second branch operates on the original SAI array with the same operation to extract local disparity. The two disparities are concatenated along the channel dimension, reduced to the channel number by a final $3\times3\times3$ convolution and LeakyReLU, and output as the super-resolved SAI array. The whole procedure can be written as:

$$F_{\mathrm{SR}} = H_{\mathrm{R}}(F_{\mathrm{HR}}) \tag{4}$$

where $H_{\mathrm{R}}(\cdot)$ denotes the re-optimization operation, F_{SR} represents the output of the SR LF image.

3. Experiment results and analysis

3.1. Datasets and experiment details

This paper trains and evaluates the proposed method on five public datasets, which are EPFL [17], HCInew [18], HCIold [19], INRIA [20], and STFgantry [21]. Detailed information is listed in Table 1. All 9×9 angular resolution LF images use central 5×5 views for training and testing. During training, every SAI is cropped into 64×64 patches with a stride of 32 to serve as ground truth. These patches are then down-sampled by bicubic interpolation to 32×32 (2× task) or 16×16 (4× task) to generate the low-resolution inputs. Data augmentation (random horizontal and vertical flips and

T a b l e 1. The five public datasets adopted in the experiments.

Datasets	EPFL [17]	HCInew [18]	HCIold [19]	INRIA [20]	STFgantry [21]	Total
Train	70	20	10	35	9	144
Test	10	4	2	5	2	23

90° rotations) increases the training set by a factor of eight. Implemented in PyTorch on an NVIDIA RTX 3090, the network is trained for 50 epochs with batch size 8, optimized by ADAM under LI loss, and the initial learning rate is halved every 15 epochs.

For quantitative evaluation, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are computed on the Y channel of the super-resolved images [12-14]. The score for each scene is obtained by averaging the metrics over all its super-resolved SAIs, the final score for each dataset is then the average across all scenes in that dataset.

3.2. Network performance evaluation

To evaluate the proposed LF SR network, 2× and 4× SR is performed on twenty-three scenes from five public datasets. Qualitative and quantitative analyses include subjective visual presentation, PSNR and SSIM calculation, and algorithm complexity, with comparisons to other methods to confirm its superiority.

Qualitative evaluation. We demonstrate the superiority of the proposed method, against eight representative LF SR algorithms (ResLF [22], LFSSR [9], LF-InterNet [12], LF-DFnet [23], LF-IInet [14], DPT [11], DistgSSR [13], LF-DGnet [15]) and the bicubic interpolation baseline. Qualitative comparisons of the super-resolved images are shown in Figs. 3 and 4.

Figure 3 compares 2× SR results on the "ISO_Chart" scene from the EPFL dataset [17] and the "MessyDesk" scene from the INRIA dataset [20]. Zooming in on red-boxed regions reveals superior restoration quality such as clear vertical textures of the trumpet-shaped object in "ISO_Chart"'s elliptical area, and high-clarity, low-artifact data cable outlines in "MessyDesk"'s elliptical area, proving better 2× LF SR image quality. Figure 4 compares 4× SR results on the "Bicycle" scene from HCInew [18] and the "Tarot Cards" scene from STFgantry [21]. Zooming into red-boxed regions shows the method better preserves intricate textures and sharpens boundaries with clean ornamental contours in "Bicycle"'s elliptical crop, and a crisper inverted human figure in "Tarot Cards"'s corresponding area. These results confirm that the network outperforms competitors in fine texture and edge fidelity, achieving superior 4× LF SR reconstruction quality. In summary, bicubic interpolation yields the worst results because it is the most rudimentary traditional algorithm. ResLF [22], LFSSR [9] and LF-DFnet [23] improve super-resolution by exploiting both spatial and angular information. LF-Internet [12] and LF-IInet [14] further boost performance through interactive frameworks. DPT [11] introduces a transformer architecture that better preserves fine details, producing higher-quality images. Our proposed method fully excavates spatial and angular features of the light field via an iterative split-fusion-and-enhancement structure, generating high-resolution images that are even closer to the ground truth.

Quantitative assessment. We further validate the SR performance of the proposed method, conducting quantitative comparisons with representative LF SR algorithms. The PSNR and SSIM results are reported in Table 2, where the best value in each column is highlighted in blue bold, and the second-best in black bold. For 2× SR, average PSNR

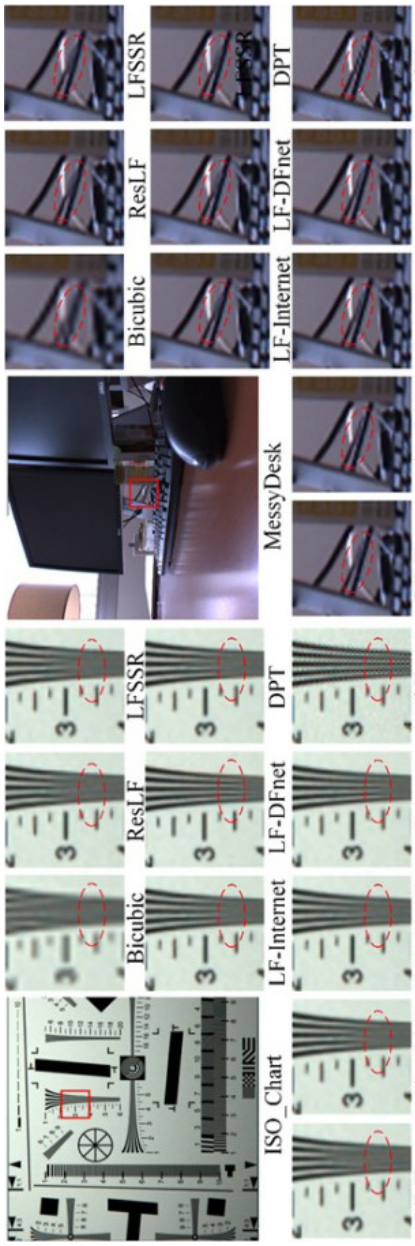

Fig. 3. Comparison of 2× SR results of different algorithms in "ISO_Chart" and "MessyDesk" scenarios.

Fig. 4. Comparison of 4× SR results of different algorithms in "Bicycle" and "Tarot Cards" scenarios.

T a b l e 2. Quantitative comparison of different methods under 2× and 4× SR tasks.

		Datasets					Average
	Method	EPFL	HCInew	HCIold	INRIA	STFgantry	PSNR/SSIM
2×	Bicubic	29.74/0.9376	31.89/0.9356	37.69/0.9785	31.33/0.9577	31.06/0.9498	32.34/0.9518
	ResLF [22]	32.75/0.9672	36.07/0.9715	42.61/0.9922	34.57/0.9784	36.89/0.9873	36.58/0.9793
	LFSSR [9]	33.69/0.9748	36.86/0.9753	43.75/0.9939	35.27/0.9834	38.07/0.9902	37.53/0.9835
	LF-Internet [12]	34.14/0.9761	37.28/0.9769	44.45/0.9945	35.80/0.9846	38.72/0.9916	38.08/0.9847
	LF-DFnet [23]	34.44/0.9766	37.44/0.9786	44.23/0.9943	36.36/0.9841	39.61/0.9935	38.42/0.9854
	DPT [11]	34.48/0.9759	37.35/0.9770	44.31/0.9943	36.40/0.9843	39.52/0.9928	38.41/0.9849
	LF-IInet [14]	34.68/0.9771	37.74/0.9789	44.84/0.9948	36.57/0.9853	39.86/0.9935	38.74/0.9859
	DistgSSR [13]	**34.91/0.9787**	**37.96/0.9796**	**44.94/0.9949**	**36.59/0.9859**	**40.40/0.9942**	**38.94/0.9866**
	LF-DGnet [15]	**34.76**/0.9771	**37.91/0.9794**	**44.92/0.9948**	**36.68**/0.9854	40.37/**0.9941**	**38.93**/0.9862
	Ours	34.75/**0.9782**	37.82/0.9791	44.85/**0.9949**	36.50/**0.9856**	**40.45/0.9941**	38.88/**0.9864**
4×	Bicubic	25.14/0.8324	27.61/0.8517	32.42/0.9344	26.82/0.8867	25.93/0.8452	27.58/0.8661
	ResLF [22]	27.46/0.8899	29.92/0.9011	36.12/0.9651	29.64/0.9339	28.99/0.9214	30.43/0.9223
	LFSSR [9]	28.27/0.9080	30.72/0.9124	36.70/0.9690	30.31/0.9446	30.15/0.9385	31.23/0.9345
	LF-Internet [12]	28.67/0.9143	30.98/0.9165	37.11/0.9715	30.64/0.9486	30.53/0.9426	31.59/0.9387
	LF-DFnet [23]	28.77/0.9165	31.23/0.9196	37.32/0.9718	30.83/0.9503	31.15/0.9494	31.86/0.9415
	DPT [11]	28.93/0.9167	31.19/0.9186	37.39/0.9720	30.96/0.9502	31.14/0.9487	31.92/0.9412
	LF-IInet [14]	**29.11**/0.9194	31.36/0.9211	**37.62/0.9737**	**31.08**/0.9516	31.21/0.9495	32.08/0.9431
	DistgSSR [13]	28.99/**0.9195**	31.38/**0.9217**	37.56/**0.9732**	30.99/0.9519	**31.65/0.9535**	**32.11/0.9440**
	LF-DGnet [15]	29.00/0.9182	**31.43/0.9217**	37.57/0.9730	31.03/**0.9520**	31.35/0.9514	32.08/0.9433
	Ours	**29.22/0.9218**	**31.47/0.9226**	**37.69/0.9737**	**31.33/0.9530**	**31.66/0.9537**	**32.27/0.9450**

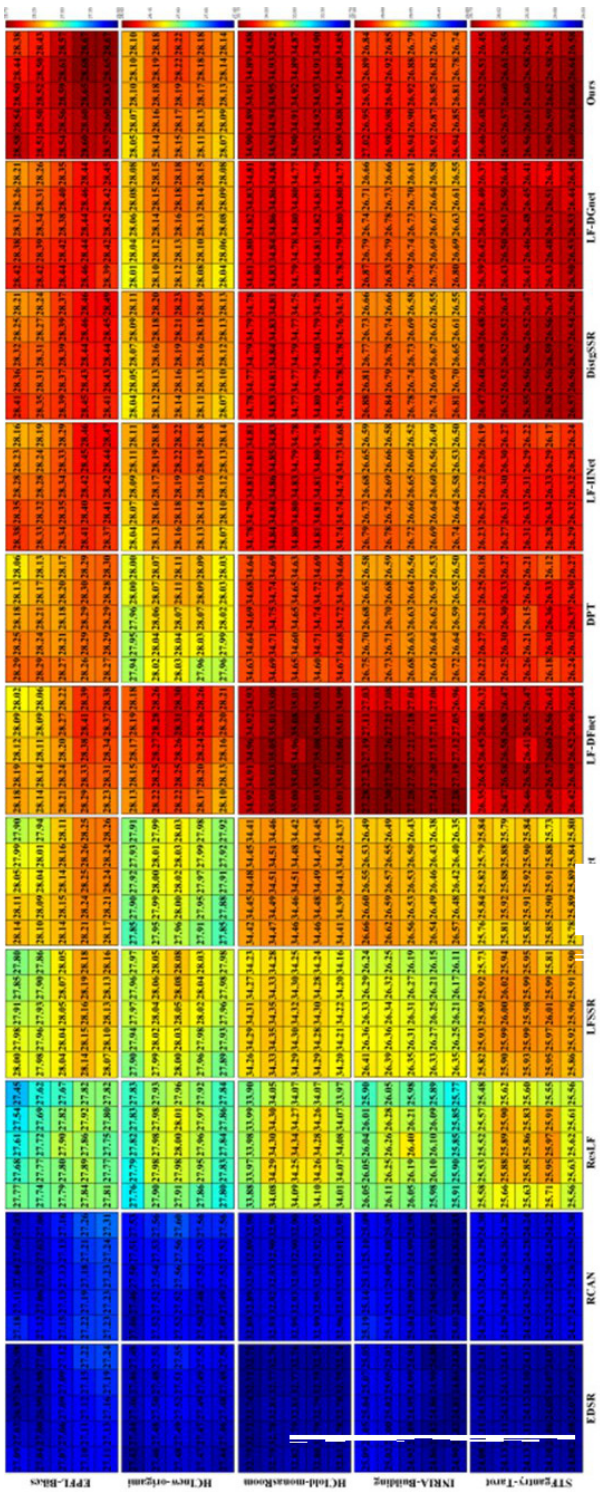

Fig. 5. PSNR distribution across angular sub-aperture views under 4× super-resolution.

is slightly lower than LF-DGnet and DistgSSR but outperforms others. SSIM ranks second only to DistgSSR. For 4× SR, achieves the highest scores across all five datasets, outperforming DistgSSR by 0.16 dB (PSNR) and 0.0010 (SSIM), and LF-DGnet by 0.19 dB (PSNR) and 0.0017(SSIM). These results confirm the network's superior reconstruction quality.

Super-resolution SAIs PSNR distribution. To evaluate performance across different angular positions, five representative scenes are selected from the test sets of the five datasets, such as EPFL-Bikes, HCInew-Origami, HCIold-MonasRoom, INRIA -Building and STFgantry-Tarot. Taking 4× SR as an example, the PSNR of each central 5×5 SAI is computed and visualized in Fig. 5, where higher PSNR values are rendered in redder tones. EDSR [24] and RCAN [25] are also included for comparison. Across all scenes, the proposed method consistently produces the reddest (highest-PSNR) regions, clearly demonstrating its superior reconstruction quality at every angular view.

Evaluation of accuracy and complexity. We intuitively demonstrate the proposed method's superiority, drawing bubble charts that jointly show PSNR, SSIM and parameter count. The bubble position indicates accuracy (the closer to the top-right, the higher PSNR/SSIM), while the bubble size represents model complexity (smaller bubbles mean fewer parameters). As shown in Fig. 6, the proposed method lies in the upper-right region with a relatively small bubble, confirming competitive SR quality with markedly fewer parameters.

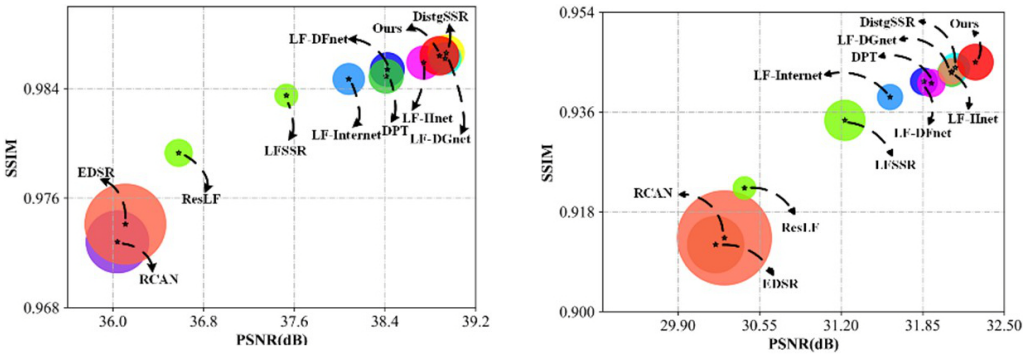

Fig. 6. Accuracy and complexity comparison of different methods under 2× and 4× SR tasks.

Algorithm complexity. To comprehensively evaluate the proposed method, it is compared with representative LF SR algorithms in parameter count (#Params), computational cost (FLOPs) and running time, and the results are reported in Table 3. In #Params, it is only higher than DistgSSR, DPT, and LFSSR but significantly lower than others. Regarding FLOPs and running time, it shows slightly larger values—attributed to MSFM's multi-angle disparity extraction via 3D convolutions, cyclic feature separation-mixing across stacked MSFMs with equal-step residual connections, and the terminal GLFR module (3D convolution-based global/local feature extraction). For accuracy, 2× SR achieves PSNR slightly lower than LF-DGnet, DistgSSR and SSIM

Table 3. The algorithm complexity comparison of the different methods.

Method	2×				4×			
	#Params/M	FLOPs/G	Runtime	PSNR/SSIM	#Params/M	FLOPs/G	Runtime	PSNR/SSIM
ResLF [22]	6.35	**37.06**	2.95	36.58/0.9793	6.79	**39.70**	0.95	30.43/0.9223
LFSSR [9]	**0.81**	**25.70**	**0.21**	37.53/0.9835	**1.61**	128.44	**0.16**	31.23/0.9345
LF-Internet [12]	4.80	47.46	1.45	38.08/0.9847	5.23	**50.10**	0.47	31.59/0.9387
LF-DFnet [23]	3.94	57.22	1.73	38.42/0.9854	3.99	57.31	0.55	31.86/0.9415
DPT [11]	3.73	57.44	7.56	38.41/0.9849	3.78	58.64	2.06	31.92/0.9412
LF-IInet [14]	4.84	56.16	0.55	38.74/0.9859	4.89	57.42	0.21	32.08/0.9431
DistgSSR [13]	3.53	64.11	1.55	**38.94/0.9866**	3.58	65.41	0.46	**32.11/0.9440**
LF-DGnet [15]	4.65	48.86	2.73	**38.93**/0.9862	4.70	50.20	0.76	32.08/0.9433
Ours	3.78	75.41	5.48	38.88/**0.9864**	3.83	100.13	2.38	**32.27/0.9450**

Table 4. Validation of the effectiveness of each network component.

Model	#Params/M	Datasets					Average PSNR/SSIM
		EPFL	HCInew	HCIold	INRIA	STFgantry	
ResBlock → MSM	3.86	29.17/**0.9214**	**31.53/0.9233**	**37.73/0.9741**	31.12/**0.9529**	**31.66/0.9540**	**32.24/0.9453**
Distgblock→MSFM	3.83	29.13/0.9204	31.41/0.9214	37.68/0.9735	31.07/0.9520	**31.53**/0.9519	32.16/0.9438
MSFM (w/o spatial)	**2.59**	29.04/0.9191	31.30/0.9203	37.62/0.9733	31.05/0.9512	31.07/0.9481	32.02/0.9424
MSFM (w/o angle)	3.31	**29.18**/0.9205	31.42/0.9219	37.67/0.9736	**31.19**/0.9524	31.50/0.9526	32.19/0.9442
MSFM (w/o EPI)	**2.87**	29.14/0.9198	31.36/0.9210	37.52/0.9732	31.09/0.9520	31.38/0.9514	32.10/0.9435
w/o GLFR	3.73	29.05/0.9198	31.36/0.9209	37.56/0.9730	31.03/0.9511	31.36/0.9508	32.07/0.9431
Ours	3.83	**29.22/0.9218**	**31.47/0.9226**	**37.69/0.9737**	**31.33/0.9530**	**31.66/0.9537**	**32.27/0.9450**

marginally inferior to DistgSSR, but outperforms other competitors. 4× SR yields PSNR/SSIM significantly superior to all. The proposed method sets the number of iterations to 16, which is determined by comparative experiments to balance the model's complexity and super-resolution accuracy. In summary, the method effectively reduces parameters and improves 4× SR performance while reasonably controlling computational complexity and running time, achieving a good balance between efficiency and performance.

3.3. Ablation experiment

Ablation experiments verify the effectiveness of each proposed component and the rationality of MSFM design, and the results are reported in Table 4. For the sub-aperture feature extraction, replacing MSM with a vanilla ResBlock (first *vs*. last rows) shows that MSM increases PSNR by 0.03 dB with a slight parameter increase, confirming its superiority. In multi-dimensional disparity mining, substituting MSFM with DistgBlock (second *vs.* last rows) yields higher PSNR/SSIM that demonstrate MSFM's advantage, while testing three MSFM variants (removing spatial/angular/EPI branches, third-fifth *vs.* last rows) reveals all variants degrade performance, proving the multi-branch design is essential. For feature re-optimization, removing GLFR (sixth *vs.* last rows) noticeably reduces PSNR/SSIM, validating the refinement module's necessity. Collectively, these experiments confirm that all proposed modules positively contribute to network's performance.

4. Conclusion

To enhance the spatial resolution of LF images, this paper proposes a multi-dimensional disparity mining SR network. Sub-aperture features are thoroughly exploited by ResAspp and MSM, while the iterative structure MSFM enriches disparity representations. An upsampling module then boosts spatial resolution, and a final global –local refinement stage further enhances the high-resolution output. In 4× SR the network achieves 32.27 dB PSNR and 0.9450 SSIM on average across five public LF datasets, delivering superior image quality. Comparative experiments confirm that the proposed approach outperforms state-of-the-art methods in both visual fidelity and objective metrics.

Acknowledgements
This work was supported by the National Natural Science Foundation of China under Grant 61601318, in part by the Patent Conversion Program of Shanxi Province under Grant No.202405009, and in part by the Fundamental Research Program of Shanxi Province under Grant No.202103021224278.

References

[1] LIU M.D., WU R.F., LUO Z.C., ZHEN J.R., ZHANG H.Q., LUO J.X., YAN L.S., WU Y.X., *Fast digital refocusing Fourier ptychographic microscopy method based on convolutional neural network*, Optics Express, **32**(1), 2024: 339–354. https://doi.org/10.1364/OE.512330

[2] Viganò N., Gil P.M., Herzog C., de la Rochefoucauld O., van Liere R., Batenburg K.J., *Advanced light-field refocusing through tomographic modeling of the photographed scene*, Optics Express, **27**(6), 2019: 7834–7856. https://doi.org/10.1364/OE.27.007834

[3] Wu G.C., Masia B., Jarabo A., Zhang Y., Wang L., Dai Q., Chai T., Liu Y., *Light field image processing: An overview*, IEEE Journal of Selected Topics in Signal Processing, **11**(7), 2017: 926–954. https://doi.org/10.1109/JSTSP.2017.2747126

[4] Wu Y.C., Wang Y.M., Liang J., Bajić I.V., Wang A.H., *Light field all-in-focus image fusion based on spatially-guided angular information*, Journal of Visual Communication and Image Representation, **72**, 2020: 102878. https://doi.org/10.1016/j.jvcir.2020.102878

[5] Yingchun Wu, Yumei Wang, Anhong Wang, Xianling Zhao, *Light field all-in-focus image fusion based on edge enhanced guided filtering*, Journal of Electronics & Information Technology, **42**(9), 2020: 2293–2301. (in Chinese)

[6] Chao W.T., Duan F.Q., Wang X.C., Wang Y.Q., Lu K., Wang G.H., *OccCasNet: Occlusion-aware cascade cost volume for light field depth estimation*, IEEE Transactions on Computational Imaging, **10**, 2024: 1680–1691. https://doi.org/10.1109/TCI.2024.3488563

[7] Chao W.T., Wang X.C., Wang Y.Q., Wang G.H., Duan F.Q., *Learning sub-pixel disparity distribution for light field depth estimation*, IEEE Transactions on Computational Imaging, **9**, 2023: 1126–1138. https://doi.org/10.1109/TCI.2023.3336184

[8] Wang Y.Q., Wang L.G., Liang Z.Y., Yang J.G., An W., Guo Y.L., *Occlusion-aware cost constructor for light field depth estimation*, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022: 19777–19786. https://doi.org/10.1109/CVPR52688.2022.01919

[9] Yeung H.W.F., Hou J.H., Chen X.M., Chen J., Chen Z.B., Chung Y.Y., *Light field spatial super-resolution using deep efficient spatial-angular separable convolution*, IEEE Transactions on Image Processing, **28**(5), 2019: 2319–2330. https://doi.org/10.1109/TIP.2018.2885236

[10] Liang Z.Y., Wang Y.Q., Wang L.G., Yang J.G., Zhou S.L., *Angular-flexible network for light field image super-resolution*, Electronics Letters, **57**(24), 2021: 921–924. https://doi.org/10.1049/ell2.12312

[11] Wang S.Z., Zhou T.F., Lu Y., Di H.J., *Detail-preserving transformer for light field image super-resolution*, Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, **36**(3), 2022: 2522–2530.

[12] Wang Y.Q., Wang L.G., Yang J.G., An W., Yu J.Y., Guo Y.L., *Spatial-angular interaction for light field image super-resolution*, [In] Vedaldi A., Bischof H., Brox T., Frahm JM. [Eds.] *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, Vol. 12368, Springer, Cham, 2020: 290–308. https://doi.org/10.1007/978-3-030-58592-1_18

[13] Wang Y.Q., Wang L.G., Wu G.C., Yang J.G., An W., Yu J.Y., Guo Y.L., *Disentangling light fields for super-resolution and disparity estimation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **45**(1), 2023: 425–443. https://doi.org/10.1109/TPAMI.2022.3152488

[14] Liu G.S., Yue H.J., Wu J.M., Yang J.Y., *Intra-inter view interaction network for light field image super-resolution*, IEEE Transactions on Multimedia, **25**, 2021: 256–266. https://doi.org/10.1109/TMM.2021.3124385

[15] Liu G.S., Yue H.J., Li K., Yang J.Y., *Disparity-guided light field image super-resolution via feature modulation and recalibration*, IEEE Transactions on Broadcasting, **69**(3), 2023: 740–752. https://doi.org/10.1109/TBC.2023.3284408

[16] Tao M.W., Hadap S., Malik J., Ramamoorthi R., *Depth from combining defocus and correspondence using light-field cameras*, 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 2013: 673–680. https://doi.org/10.1109/ICCV.2013.89

[17] Rerabek M., Ebrahimi T., *New light field image dataset*, Proceedings of the 8th International Conference on Quality of Multimedia Experience, Lisbon, Portugal, 2016.

[18] Honauer K., Johannsen O., Kondermann D., Goldluecke B., *A dataset and evaluation methodology for depth estimation on 4D light fields*, [In] Lai S.H., Lepetit V., Nishino K., Sato Y. [Eds.],

Computer Vision – ACCV 2016. Lecture Notes in Computer Science, Vol. 10113, Springer, Cham, 2017: 19–34. https://doi.org/10.1007/978-3-319-54187-7_2

[19] WANNER S., MEISTER S., GOLDLUECKE B., *Datasets and benchmarks for densely sampled 4D light fields*, Vision, Modeling, and Visualization. Switzerland: Computer Graphics Forum, **13**, 2013: 225–226.

[20] LE PENDU M., JIANG X.R., GUILLEMOT C., *Light field inpainting propagation via low rank matrix completion*, IEEE Transactions on Image Processing, **27**(4), 2018: 1981–1993. https://doi.org/10.1109/TIP.2018.2791864

[21] VAISH V., ADAMS A., The (new) stanford light field archive [EB/OL], 2024: 3–15.

[22] ZHANG S., LIN Y.F., SHENG H., *Residual networks for light field image super-resolution*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 11038–11047. https://doi.org/10.1109/CVPR.2019.01130

[23] WANG Y.Q., YANG J.G., WANG L.G., YING X.Y., WU T.H., AN W., GUOY.L., *Light field image super-resolution using deformable convolution*, IEEE Transactions on Image Processing, **30**, 2020: 1057–1071. https://doi.org/10.1109/TIP.2020.3042059

[24] LIM B., SON S., KIM H., NAH S., LEE K.M., *Enhanced deep residual networks for single image super-resolution*, 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, USA, 2017: 1132–1140. https://doi.org/10.1109/CVPRW.2017.151

[25] ZHANG Y.L., LI K.P., LI K., WANG L.C., ZHONG B., FU Y., *Image super-resolution using very deep residual channel attention networks*, [In] FERRARI V., HEBERT M., SMINCHISESCU C., WEISS Y. [Eds.] *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, Vol. 11211, Springer, Cham, 2018: 294–310. https://doi.org/10.1007/978-3-030-01234-2_18

Received November 11, 2025
in revised form December 17, 2025