LGFN: Lightweight Light Field Image Super-Resolution using Local Convolution Modulation and Global Attention Feature Extraction

Zhongxin Yu, Liang Chen[*], Zhiyun Zeng, Kunping Yang, Shaofei Luo, Shaorui Chen, Cheng Zhong
Fujian Normal University
wuyizhizi555@163.com   cl_0827@126.com   1304370458@qq.com
kunpingyang@fjnu.edu.cn   {shaofeiluo,1589177538,2998233739}@qq.com

Abstract

Capturing different intensity and directions of light rays at the same scene, Light field (LF) can encode the 3D scene cues into a 4D LF image, which has a wide range of applications (i.e., post-capture refocusing and depth sensing). LF image super-resolution (SR) aims to improve the image resolution limited by the performance of LF camera sensor. Although existing methods have achieved promising results, the practical application of these models is limited because they are not lightweight enough. In this paper, we propose a lightweight model named LGFN, which integrates the local and global features of different views and the features of different channels for LF image SR. Specifically, owing to neighboring regions of the same pixel position in different sub-aperture images exhibit similar structural relationships, we design a lightweight CNN-based feature extraction module (namely, DGCE) to extract local features better through feature modulation. Meanwhile, as the position beyond the boundaries in the LF image presents a large disparity, we propose an efficient spatial attention module (namely, ESAM) which uses decomposable large-kernel convolution to obtain an enlarged receptive field and an efficient channel attention module (namely, ECAM). Compared with the existing LF image SR models with large parameter, our model has a parameter of 0.45M and a FLOPs of 19.33G, which has achieved a competitive effect. Extensive experiments with ablation studies demonstrate the effectiveness of our proposed method, which ranked the second place in the Track 2 Fidelity & Efficiency of NTIRE2024 Light Field Super Resolution Challenge and the seventh place in the Track 1 Fidelity.

1. Introduction

LF cameras can capture varying intensities and directions of light rays within the same scene, encoding the 3D scene cues into a 4D LF image (comprising spatial and angular dimensions). This technology finds wide applications, including post-capture refocusing[1, 2], depth sensing[3–5], virtual reality[6, 7], and view rendering[8–11]. However, due to the limitation of sensor performance, there exists a trade-off between the spatial resolution and angular resolution of LF images. How to improve the resolution of LF images is currently a prominent research challenge.

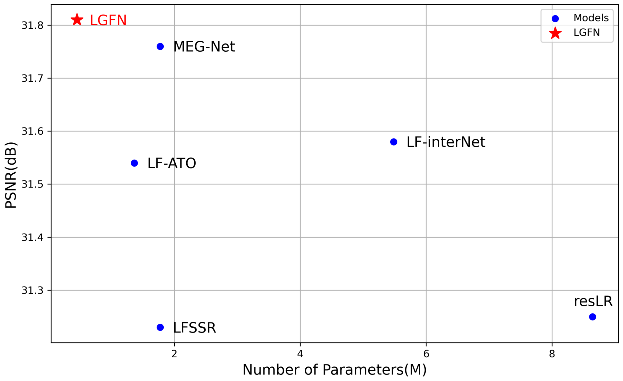

Figure 1. Comparisons of the parameters and PSNR of different LF image SR methods.

The traditional LF SR method[12–16] mainly focuses on how to find sub-pixel information and warp multi-view images based on estimated disparities. However, the performance of these methods heavily depends on accurate estimated disparities, which is difficult to achieve in low-resolution LF images and complex imaging environments such as occlusion and non-Lambert reflection[17].

In recent years, deep learning-based methods have been widely used. Yoon et al.[18] proposed the first CNN-based LF image SR model (i.e., LFCNN), which used SRCNN to super-resolve each sub-aperture image (SAI). Afterwards, many methods have adopted the CNN-based methods to integrate different angle information to improve the performance of SR[19–23]. Besides directly processing the

[*] Corresponding author

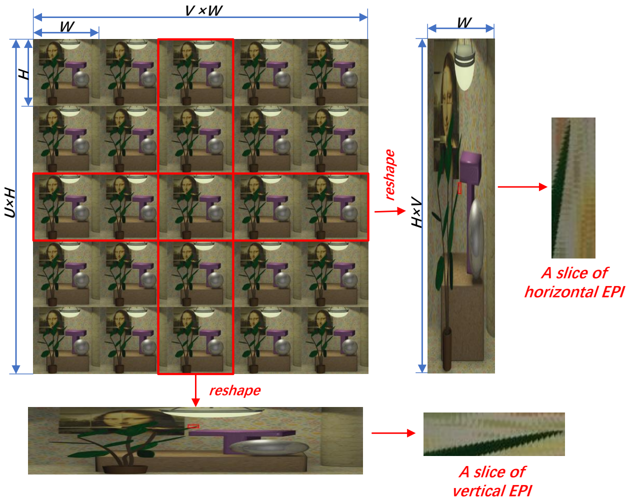

Figure 2. The epipolar plane images(EPI) sample of 4D LF is acquired with fixed angular coordinate and a fixed spatial coordinate. Specifically, the horizontal EPI is obtained with constants u and h, and the vertical EPI is obtained with constant v and w. On the one hand, the EPIs capture spatial structures such as edges or textures, and the adjacent areas corresponding to the same pixel position across different SAIs exhibit similar structural relationships. On the other hand, the EPIs reflect the disparity information via line patterns of different slopes, whereas positions located outside the boundary in the LF image exhibit a large parallax.

4D LF data, some methods extracted two kinds of features by designing spatial and angular feature extractor, and interacted with each other[24–28]. Apart from the CNN-based LF image SR methods, Transformer-based LF methods have also been proposed. Wang et al.[26] proposed a detail-preserving Transformer (DPT) for LF image SR. Liang et al.[29] proposed a simple yet efficient Transformer method for LF image SR. Liang et al.[30] proposed EPIT to LF image SR by learning non-local space and angle cooperation. Jin et al.[31] proposed DistgEPIT model that learn global features and local features of LF images by designing an attention branch and a convolution branch respectively. While existing methods have achieved promising results, the practical application of these models is limited due to excessive parameters and FLOPs. As shown in Fig.1, the parameters of some existing LF image SR methods are mostly above 1M. This limitation prompts our research into lightweight LF image SR.

As shown in Fig.2, adjacent areas at the same pixel position across different SAIs exhibit similar structural relations, which are suitable for processing by the local feature extraction module. On the other hand, the position outside the boundary in the LF image exhibits a large parallax, which requires aggregation of contextual features across SAIs for processing.

To consider these two aspects and the requirement of lightweight model, we propose a lightweight local and global feature learning model named LGFN. By integrating both local and global features of different views and the features of different channels, our lightweight model can achieve competitive results compared with the existing model with larger parameters. Specifically, our model chooses the convolution module with local representation. Different from the existing CNN-based methods[19–23] which use complex network structure, we propose a simple yet efficient convolution module designed to extract local features through feature modulation performed by two parallel convolution branches.

In addition, we choose attention mechanism to extract contextual features.Different from the existing Transformer-based methods [26, 29–31]with quadratic complexity over the number of visual tokens, we propose a simple yet efficient spatial attention module, whose attention weight branch uses decomposable large-kernel convolution to obtain an enlarged receptive field, and multiplies it with identity branch to extract contextual features. Besides, an efficient channel attention module (namely, ECAM) is introduced to enhance the features between channels. In order to further refine the feature extraction, we extract the local and global features along the horizontal and vertical directions respectively.

Our main contribution can be summarized as:

- We design a lightweight convolution modulation module named DGCE to extract the local spatial features of LF images. A lightweight spatial attention module named ESAM with enlarged receptive field is designed to extract global features. In order to further refine the feature extraction, we extract the local and global features along the horizontal and vertical directions respectively.
- We design an efficient channel attention module named ECAM and use the statistical information of channel direction to model the correlation between different channels.
- We propose a light-weight LF image SR model named LGFN, which has a parameter of 0.45M and a FLOPs of 19.33G. Compared with the existing LF image SR models with large parameter, it has achieved a competitive effect, and won the second place in the Track 2 Fidelity Efficiency of NTIRE2024 Light Field Super Resolution Challenge and the seventh place in the Track 1 Fidelity.

2. Related Work

LR image SR methods can be divided into traditional non-learning methods, CNN-based methods and Transformer-based methods.

2.1. Traditional Methods

The traditional LF image SR methods mainly focuses on how to find sub-pixel information and warp multi-view images based on estimated disparities. Based on estimated disparities, Bishop et al.[12] used a Bayesian deconvolution method to super-resolve LF images. Wanner et al.[13] used EPI to estimated disparity maps and proposed variational framework for LF image SR. Farrugia et al.[14] proposed an example-based LF image SR method, enhancing spatial resolution consistently across SAIs through learning linear projections from reduced-dimension subspaces and angular super-resolution via multivariate ridge regression. Besides, optimization-based methods have also been proposed. Alain et al.[16] adopted an optimization method to solve ill-posed LF image SR problem based on sparsity prior. Rossi et al.[15] coupled the multi-frame information with a graph regularization, adopted convex optimization method to solve LF image SR problem.

However, the performance of these methods depends heavily on accurate estimated disparities, it is difficult to achieve in low-resolution LF images and complex imaging environments such as non-Lambertian surfaces or occlusions[17].

2.2. CNN-based Methods

In recent years, deep learning-based method have been widely used. Yoon et al. [18] proposed the first CNN-based LF image SR model (i.e., LFCNN), which used SRCNN to super-resolve each SAI. Similarly, Yuan et al.[32] uses EDSR to super-resolve each SAI. Afterwards, many methods have adopted the CNN-based methods to integrate different angle information to improve the performance of SR. Wang et al.[19]proposed a bidirectional recurrent CNN network iteratively model spatial relations between horizontally or vertically adjacent SAIs. Zhang et al.[20] proposed resLF network that used four-branch residual network extracted features from SAI images along four different angular directions. Zhang et al.[23] proposed a 3D convolutions network extracted features from SAI images along different angular directions. Cheng et al. [21]considered the characteristics of internal similarity and external similarity of LR images, and fused these two complementary features for LF image SR. Meng et al. [17] directly used 4D convolution to extract the angle information and spatial information of the LF image. Wang et al.[22]designed an angular deformable alignment module (ADAM) for feature-level alignment, and proposed a collect-and-distribute approach to perform bidirectional alignment between the center-view feature and each side-view feature. In addition to directly processing the 4D LF data, some methods disentangled the 4D LFs into different subspaces for SR. Wang et al. [25] proposed a spatial and angular feature extractor to extract the corresponding spatial and angular information from the MacPI image, and proposed LF-InterNet[27]and DistgSSR[28]to repetitively interact the two features.

Besides the aforementioned methods to improve SR performance, some methods try to solve the complex degradation problem facing the real world. To address the issue of the domain gap in LF image SR, Cheng et al.[33] proposed a ’zero-shot’ learning framework. They divided the end-to-end model training task into three sub-tasks: pre-upsampling, view alignment, and multi-view aggregation, and subsequently tackled each of these tasks separately by using simple yet efficient CNN networks. Xiao et al.[34] proposed the first real-world LF image SR dataset called LytroZoom, and proposed an omni-frequency projection network(OFPNet), which deals with the spatially variant degradation by dividing features into different frequency components and iteratively enhancing them. Wang et al.[35] developed a LF degradation model based on the camera imaging process, and proposed LF-DMnet that can modulate the degradation priors into CNN-based SR process.

2.3. Transformer-based Methods

In addition to the CNN-based LF image SR methods, Transformer-based LF methods have also been proposed. Wang et al.[26] proposed a detail-preserving Transformer (DPT) for LF image SR, which regards SAIs of each vertical or horizontal angular view as a sequence, and establishes long-range geometric dependencies within each sequence via a spatial-angular locally-enhanced self-attention layer. Liang et al. [29]proposed a simple yet efficient Transformer method for LF image SR, in which an angular Transformer is designed to incorporate complementary information among different views, and a spatial Transformer is developed to capture both local and long-range dependencies within each SAI. By designing three granularity aggregation units to learn LF feature, Wang et al.[36]proposed a multi-granularity aggregation Transformer (MAT) for LF image SR; Liang et al. [30] proposed EPIT to LF image SR by learning non-local space and angle cooperation. Jin et al. [31] proposed DistgEPIT model that learns global features and local features of LF images by designing an attention branch and a convolution branch respectively.

Although the existing models have achieved promising results, their model parameters and FLOPs are not lightweight enough, which limits their practical application. In order to solve these problems, we propose a lightweight SR model of LF image by designing efficient modules.

3. Method

As mentioned above, the LF image SR needs to consider the local similarity between SAI subgraphs on the one hand, and the disparity problem behind different subgraphs on the other hand, which urges us to consider the methods of local and global feature extraction.

In order to design a lightweight model with fewer parameters and FLOPs, we choose to reduce the high-dimensional feature space to the low-dimensional feature subspace, and design an efficient local and global feature extraction model to achieve LF image SR.

3.1. Network Architecture

Specifically, as illustrated in Fig.3, our LF image SR model mainly consists of three components: shallow feature extraction, deep feature extraction and up-sampling module. Given an input LF low-resolution image $F_{LR} \in R^{U \times V \times H \times W}$ denote an LR SAI array with $U \times V$ SAIs of resolution $H \times W$. Our method takes F_{LR} as its input and generates a HR SAI array of size $F_{HR} \in R^{U \times V \times sH \times sW}$,where s denotes the upsampling factor.

Firstly, in the shallow feature extraction part, the low-resolution 4D LF image is upsampled using bilinear interpolation to the size of $sH \times sW$. Meanwhile, it is converted to $F_0 \in R^{1 \times UV \times H \times W}$format and passed through a 1×3×3 spatial convolution to extract the shallow feature F_{init}, and the number of channels is increased from 1 to 64:

$$F_{init} = H_{conv}(F_{LR}) \tag{1}$$

where $H_{conv}(.)$ denotes 3D convolution operation.

Next, the shallow features F_{init} pass through the jump connection and the deep feature extraction module(DFEM) respectively to obtain the jump connection feature and the deep feature, and they are fused by a 3D convolution process.

$$F_1 = H_{DFEM}(F_{init}) + F_{init} \tag{2}$$

$$F_{fuse} = H_{conv}(F_1) \tag{3}$$

where $H_{DFEM}(.)$ and $H_{conv}(.)$ denote deep feature extraction module and 3D convolution operation, respectively.

Finally, the fused features F_{fuse} pass through an up-sampling module consisting of 1×1 convolution, piexlshuffle, LeakyReLU and 3×3 convolution. In addition, the final restored image F_{HR} is obtained by adding the initial features after bilinear interpolation:

$$F_{HR} = H_{upsampling}(F_{fuse}) + H_{bilinear}(F_{LR}) \tag{4}$$

where $H_{upsamping}(.)$ denotes up-sampling module, and $H_{bilinear}(.)$ denotes bilinear interpolation.

3.2. Local and Global Deep Feature Extraction

DFEM includes seven local and global feature extraction modules (LGFM). The LGFM consists of three components: double-gated convolution extraction module (DGCE), efficient spatial attention module (ESAM) and efficient channel attention module (ECAM), as shown in Fig.3(a).

3.2.1 Double-Gated Convolution Extraction Module

Owing to neighboring regions of the same pixel position in different SAIs exhibit similar structural relationships, which is suitable for processing with local feature extraction module.

Some studie[37–39]indicate that modulation mechanism provides satisfactory performance and is theoretically efficient (in terms of parameters and FLOPs). Therefore, we design a local feature extraction module based on feature modulation, as shown in Fig.3(b). In order to extract the local features better, the shallow features first undergo a 1×1 convolution, and then are cut into two halves along the channel. One half of the features undergoes a 3×3 depth-wise convolution and GELU function, and the other half of the features undergoes pixel-wise multiplication with the corresponding pixels to obtain the enhanced local features. After they are added to each other, they are fused by a 1×1 convolution:

$$F_{21}, F_{22} = Split(H_{conv1}(F_{init})) \tag{5}$$

$$F_{23} = \Phi(H_{dwconv3}(F_{21})) \odot F_{22} + \Phi(H_{dwconv3}(F_{22})) \odot F_{21} \tag{6}$$

$$F_{DGCE} = H_{conv1}(F_{23}) \tag{7}$$

$$F_{24} = H_{conv1}(F_{init} + F_{DGCE}) \tag{8}$$

where $\Phi(.)$ denotes activation function GELU(.), $\odot$ denotes element-wise product, $Split(.)$ denotes split features along the channel, $H_{conv1}(.)$ and $H_{dwconv3}(.)$ denote 1×1 convolution and 3×3 depth-wise convolution respectively.

3.2.2 Efficient Spatial Attention Module

Owing to the position beyond the boundaries in the LF image presents a large disparity, which requires aggregate context features among different SAIs, therefore we propose a simple yet efficient spatial attention module, as shown in Fig.3(c). In order to reduce FLOPs, a 1×1 convolution is used to reduce the number of channels, and then strided convolution and max pooling are used to further reduce the height and width of features. In order to further increase the receptive field of spatial attention, the large-kernel convolution is decomposed into a depth-wise convolution[37], a dilated convolution and a 1×1 point convolution, which can capture long-range relationships while maintaining low computational cost and few parameters. Then, the spatial resolution is restored to the original scale by up-sampling, and the number of channels is restored to the original number by a convolution. Therefore, attention with a large receptive field is obtained, which is convenient for the next attention calculation. The difference between our ESAM and other spatial attention modules is that the receptive field has been enlarged.

$$F_{25} = H_{conv1}(F_{24}) \tag{9}$$

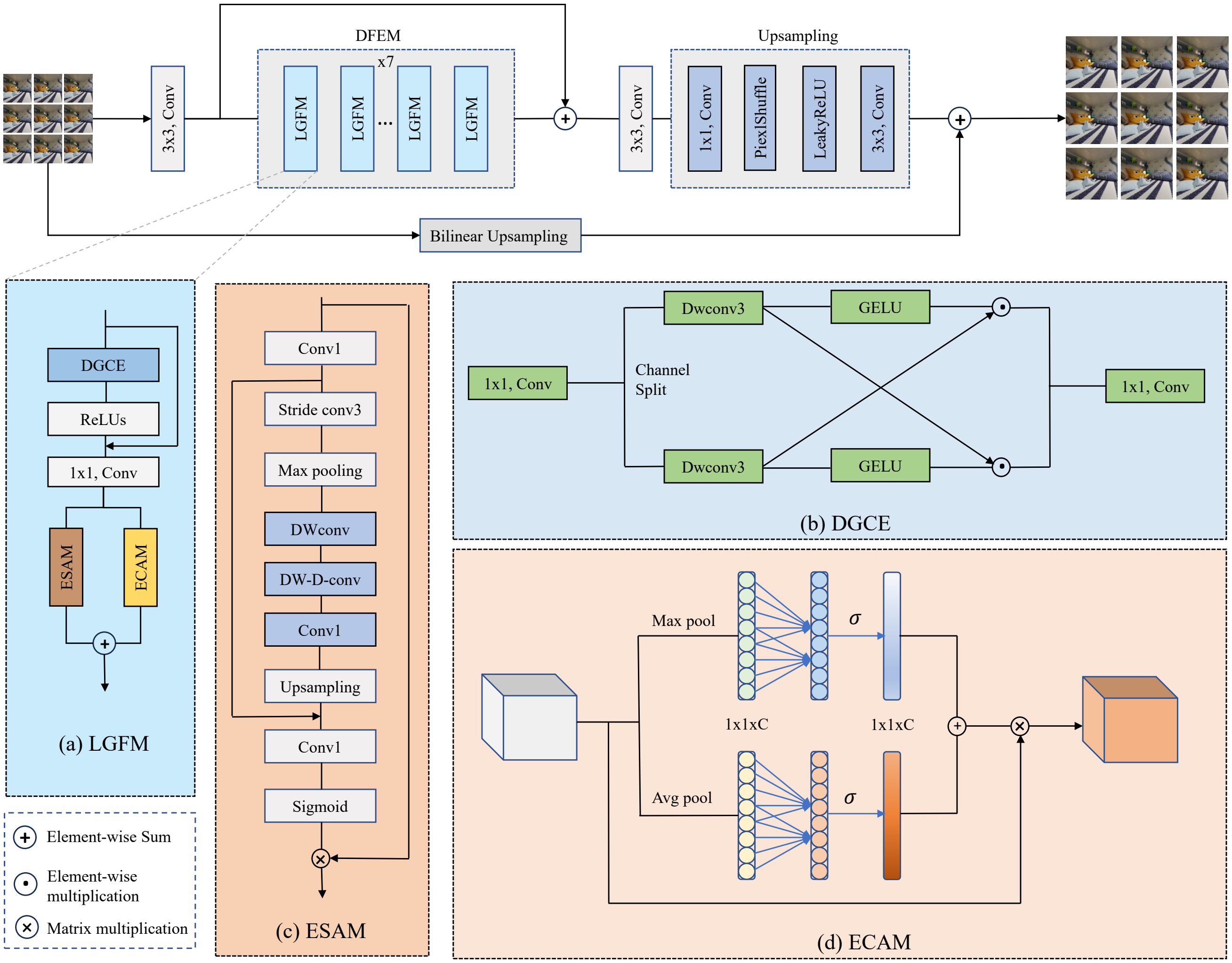

Figure 3. An overview of our LGFN network. (a) Local and global deep feature extraction module (LGFM); (b) Double-gated convolution extraction module (DGCE); (c) Efficient spatial attention module (ESAM); (d) Efficient channel attention module (ECAM). Given SAIs as inputs, we adopt bilinear upsamping to initial content of the original images. For feature extraction, we first use a 3D convolution to extract shallow features, then use the deep feature extraction module to get them, and finally use the upsampling module to obtain ultimate super-resolved SAI results. The depth feature extraction module (DFEM) includes seven local and global feature extraction modules, which are composed of DGCE, ESAM and ECAM.

$$F_{26} = H_{Maxpool}(H_{stride}(F_{25})) \tag{10}$$

$$F_{27} = H_{unsampling}(H_{LKA}(F_{26})) \tag{11}$$

$$F_{28} = H_{conv1}(F_{25} + F_{27}) \tag{12}$$

$$F_{29} = sigmoid(F_{28}) \otimes F_{24} \tag{13}$$

where $H_{LKA}(.)$ denotes decomposable large-kernel convolution operation.

3.2.3 Efficient Channel Attention Module

Some studies[20, 22, 25]show that the channel-wise features can improve LF image SR. After ESAM, we further design an efficient channel attention, as shown in Fig.3(d). For the input features, after adaptive maximum pooling and adaptive average pooling, each channel and its three adjacent channels are convolved with convolution kernel of 3 to capture local cross-channel interaction information, and two types of channel attention are obtained by sigmoid function, and then the channel attention is calculated after adding them:

$$F_{30} = H_{conv3}(H_{Maxpool}(F_{29}) \tag{14}$$

$$F_{31} = H_{conv3}(H_{Avgpool}(F_{29}) \tag{15}$$

$$F_{32} = (sigmoid(F_{30}) + sigmoid(F_{31})) \otimes F_{29} \tag{16}$$

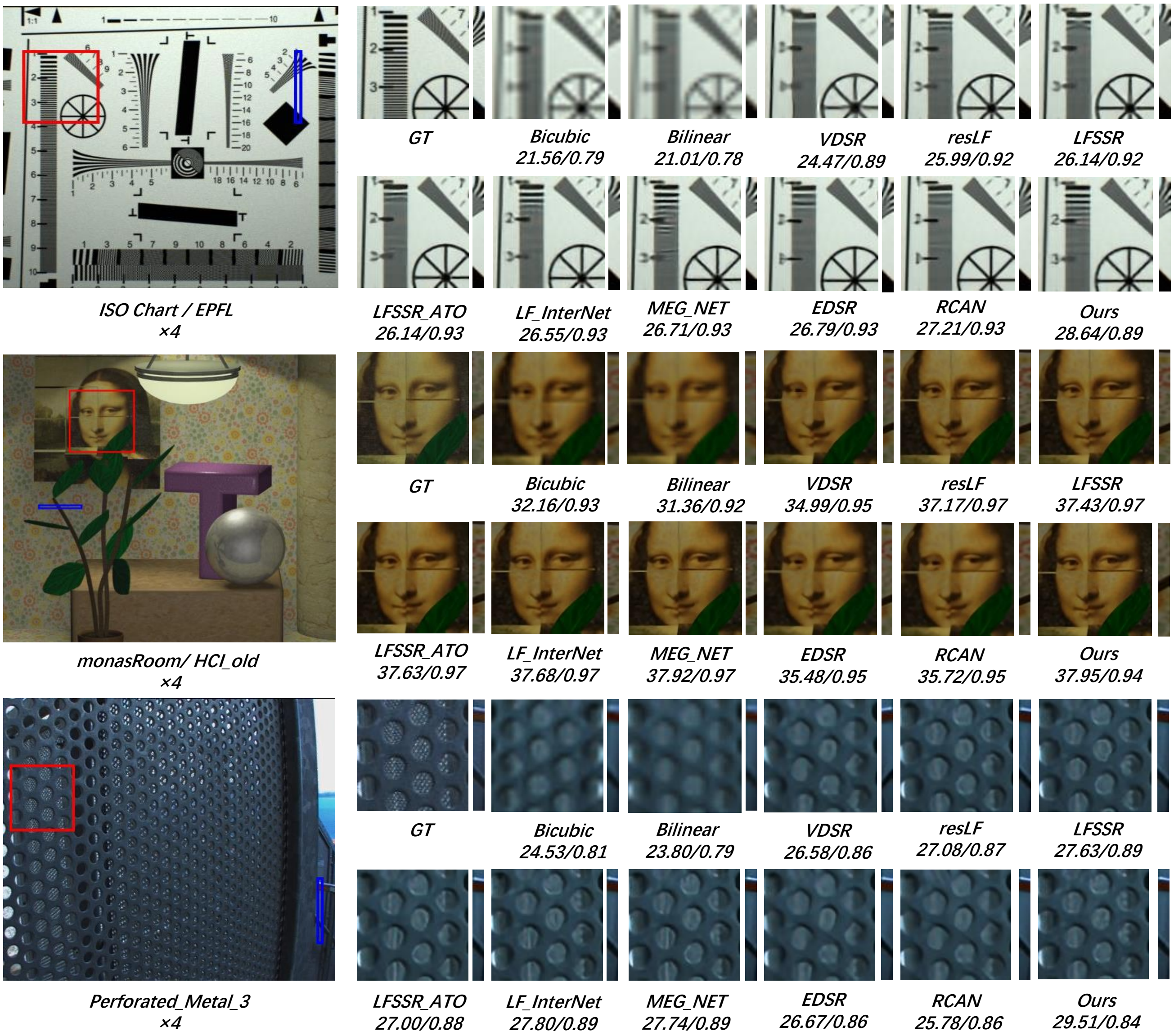

Figure 4. Qualitative results for 4x SR. The super-resolved center view images are presented for detailed texture comparison. The corresponding PSNR/SSIM scores of different methods on the presented scenes are also reported below.

In order to further refine the feature extraction, LGFM extract the local and global features along the horizontal and vertical directions respectively.

4. Experiments

In this section, we first describe the experimental details, and then carry out specific control experiments and ablation experiments.

4.1. Datasets and Implementation Details

We used the five public LF datasets: EPFL[40], HCInew [41], HCIold[42], INRIA[43], and STFgantry[44], following the same training and testing partition as in[22].

Data Augmentation. All LFs in the released datasets used the bicubic downsampling approach to generate LF patches of size 32×32. We performed random horizontal flipping, vertical flipping, and 90-degree rotation to augment the training data by 8 times. Note that, the spatial and angular dimension need to be flipped or rotated jointly to maintain LF structures.

Regularization. Our network was trained using the L1 loss and FFT Charbonnier loss with weights of 0.01 and 1 respectively. Optimized using the Adam method with $\beta1 = 0.9$, $\beta2 = 0.999$ and a batch size of 1. Our model was imple-

Table 1. Overall PSNR/SSIM metrics comparison among the other prestigious approaches for 4 x SR. The best results are in red, the second best in black.

Methods	#Param	EPFL	HCInew	HCIold	INRIA	STFganry	Average
Bilinear	-	24.57 / 0.8158	27.09 / 0.8397	31.69 / 0.9256	26.23 / 0.8757	25.20 / 0.8261	26.95 / 0.8566
Bicubic	-	25.14 / 0.8324	27.61 / 0.8517	32.42 / 0.9344	26.82 / 0.8867	25.93 / 0.8452	27.58 / 0.8701
VDSR[45]	0.665M	27.25 / 0.8777	29.31 / 0.8823	34.81 / 0.9515	29.19 / 0.9204	28.51 / 0.9009	29.81 / 0.9066
EDSR [46]	38.89M	27.84 / 0.8854	29.60 / 0.8869	35.18 / 0.9536	29.66 / 0.9257	28.70 / 0.9072	30.20 / 0.9118
RCAN [47]	15.36M	27.88 / 0.8863	29.63 / 0.8886	35.20 / 0.9548	29.76 / 0.9276	28.90 / 0.9131	30.27 / 0.9141
resLF [20]	8.646M	28.27 / 0.9035	30.73 / 0.9107	36.71 / 0.9682	30.34 / 0.9412	30.19 / 0.9372	31.25 / 0.9322
LFSSR [24]	1.774M	28.27 / 0.9118	30.72 / 0.9145	36.70 / 0.9696	30.31 / 0.9467	30.15 / 0.9426	31.23 / 0.9370
LF-ATO [48]	1.364M	28.52 / 0.9115	30.88 / 0.9135	37.00 / 0.9699	30.71 / 0.9484	30.61 / **0.9430**	31.54 / 0.9373
LF-InterNet[25]	5.483M	28.67 / **0.9162**	**30.98 / 0.9161**	**37.11 / 0.9716**	30.61 / **0.9491**	**30.53** / 0.9409	31.58 / **0.9388**
MEG-Net[23]	1.775M	28.74 / **0.9160**	**31.10 / 0.9177**	**37.27 / 0.9716**	30.66 / **0.9490**	**30.77 / 0.9453**	31.71 / **0.9399**
LGFN-C	0.45M	**30.18** / 0.8698	30.42 / 0.8370	36.31 / 0.9283	**32.05** / 0.9040	30.05 / 0.9214	**31.80** / 0.8921
LGFN-P	0.45M	**30.05** / 0.8677	30.51 / 0.8681	36.29 / 0.9282	**32.08** / 0.9037	30.11 / 0.9207	**31.81** / 0.8977

Table 2. Ablation experiments operated on 4x SSR task. Note that the mode in the table refers to the connection mode of ESAM and ECAM modules.

Mode	#Param	DGCE	ESAM	ECAM	EPFL	HCLnew	HCLold	INRIA	STFgantry	Average	PSNR
Parallel	453.6k	✓	✓	✓	30.0461	30.5145	36.2853	30.0823	30.1098	31.8076	Baseline
Cascade	453.6k	✓	✓	✓	30.1782	30.4169	36.3102	32.0481	30.0533	31.8014	-0.0062
-	452.9k	✓	✓	×	29.8148	30.5509	36.4470	31.7838	30.3020	31.7804	-0.0272
-	409.5k	✓	×	✓	29.8481	30.5490	36.1630	31.7874	30.1549	31.6995	-0.1081
-	409.5k	✓	×	×	29.8156	30.3048	36.0778	32.1215	29.9114	31.6462	-0.1614
Parallel	147.0k	×	✓	✓	27.4510	27.9710	32.7504	28.9564	26.7003	28.7796	-3.0280

mented in PyTorch on a PC with a NVidia RTX 3060 GPU. The learning rate was initially set to 2x10-4 and decreased by a factor of 0.5 for every 15 epochs. The training was stopped after 100 epochs.

We used the PSNR and SSIM computed only on the Y channel of images as quantitative metrics for performance evaluation. To compute the metric scores for a dataset containing M scenes, we firstly computed the average score of each scene by separately averaging the scores over all SAIs. Then metric score for the dataset is determined by averaging the scores over the M scenes.

4.2. Comparison to state-of-the-art methods

We compared LGFN to several state-of-the-art methods, including five SISR methods: Bilinear, Bicubic, VDSR [45], EDSR[46], RCAN[47] and other five recent LF image SR methods: resLF[49], LFSSR [32], LF-ATO[48], LF-InterNet[25], and MEG_Net[23].

Quantitative Results. As shown in Table 1, compared with other models with larger parameters, our model is very lightweight and has achieved competitive results. Specifically, the parameters of our model are the smallest, our model is only 25.35% of the parameters of MEG-Net model, but it has achieved a better average PSNR value, which shows the lightweight characteristics of our model. In addition, our model has achieved remarkable results on EPFL and INRIA datasets.

Qualitative Results. As shown in Fig.4, regarding qualitative performance, the propose LGFN has proved that it has ability to produce trustworthy details and sharp structures. For SISR methods, VDSR, EDSR and RCAN tends to produce artifacts, and the restored texture details are not clear enough. For LFSR methods, the proposed LGFN has ability to discriminate more dense details. Specifically, in figure ISO_Chart, the figure recovered by our model is clearer than other figures, and there are fewer artifacts. In the figure Perforated_Metal_3, the graph restored by our model has more material texture.

4.3. Ablation Study

In this section, we further prove the effectiveness of several core parts of the proposed LGFN model through ablation experiments.

1) Connection mode of ECAM and ESAM modules. The connection modes of ECAM and ESAM are classified into cascade connection and parallel connection. In order to verify which connection mode is more effective, we design two models: cascade and parallel, and their corresponding models are LGFN-C and LGFN-P respectively, where LGFN-C is the NTIRE2024 LF image SR competition model. As shown in Table 2, the LGFN-P is better than LGFN-C. The main model of this paper is LGFN-P.

Table 3. Our team achieved second place on the leader board (last three rows) in the NTIRE-2024 Track 2 Fidelity & Efficiency test dataset, with quantitative results of 30.05 dB PSNR (average) and 0.924 SSIM (average).

Methods	#Params	Lytro	Synthetic	Average
Bicubic	—	25.11 / 0.8404	26.46 / 0.8352	25.79 / 0.8378
VDSR [45]	0.67 M	27.05 / 0.8888	27.94 / 0.8703	27.49 / 0.8795
EDSR [46]	38.89 M	27.54 / 0.8981	28.21 / 0.8757	27.87 / 0.8869
RCAN [47]	15.36 M	27.61 / 0.9001	28.31 / 0.8773	27.96 / 0.8887
resLF [20]	8.65 M	28.66 / 0.9260	29.25 / 0.8968	28.95 / 0.9114
LFSSR [24]	1.77 M	29.03 / 0.9337	29.40 / 0.9008	29.21 / 0.9173
LF-ATO [48]	1.36 M	29.09 / 0.9354	29.40 / 0.9012	29.24 / 0.9183
LF-InterNet [25]	5.48 M	29.23 / 0.9369	29.45 / 0.9028	29.34 / 0.9198
MEG-Net [23]	1.78 M	29.20 / 0.9369	29.54 / 0.9036	29.37 / 0.9203
BITSMBU [49]	0.66 M	**30.32 / 0.9425**	**30.00 / 0.9095**	**30.16 / 0.9260**
Ours (LGFN-C)	0.45 M	**30.19 / 0.9402**	**29.92 / 0.9079**	**30.05 / 0.9240**
IIR-Lab [49]	0.83 M	29.96 / 0.9238	30.14 / 0.9407	29.96 / 0.9238

2) LGFN w/o DGCE. The DGCE module is used to extract local feature. To demonstrate the effectiveness of the DGCE module, we remove this module and use parallel ECAM and ESAM modules. As shown in Table 2, the PSNR value is decreased dramatically from 31.8076 dB to 28.7796 dB for 4x SR without DGCE module, and the drop value is 3.028dB. Experiment shows that DGCE module is effective in feature extraction.

3) LGFN w/o ESAM. The ESAM module is used to extract global LF image spatial feature. To demonstrate the effectiveness of the ESAM module, we remove this module. As shown in Table 2, the PSNR value is decreased module, the drop value is 0.1081dB.

4) LGFN w/o ECAM. The ECAM module is used to extract channel feature. To demonstrate the effectiveness of the ECAM module, we remove this module. As shown in Table 2, and the PSNR value is decreased from 31.8076 dB to 31.7804 dB for 4x SR without ECAM module, the drop value is 0.0272dB.

5) LGFN w/o ECAM and ESAM. The ESAM and ECAM module are used to extract global feature. To demonstrate the effectiveness of the ECAM and ESAM modules, we remove these modules. As shown in Table 2, the PSNR value is decreased from 31.8076 dB to 31.6462 dB for 4x SR without them, and the drop value is 0.1614dB, which proves the effectiveness of attention module.

4.4. NTIRE 2024 LFSR Challenge Results

The test set of NTIRE2024 LFSR challenge including 16 synthetic LFs and 16 real-world LFs captured by Lytro camera. As shown in Table 3, we proposed a model which ranked the second place in the Track 2 Fidelity & Efficiency of NTIRE2024 Light Field Super Resolution Challenge with 30.05dB PSNR on the LFSR test dataset.

5. Conclusion and Feature Work

In this paper, we investigated the task of lightweight LF image SR and proposed a lightweight LF image SR model named LGFN based on the local similarity and global disparity of SAIs. As a lightweight model, we proposed a feature modulation-based CNN module to extract local features efficiently. Besides, we designed an efficient spatial attention module which uses decomposable large-kernel convolution to enlarge the receptive field and an efficient channel attention module to extract the global features of the LF image. By learning local and global features, our lightweight model has achieved competitive results and ranked the second place in the Track 2 Fidelity & Efficiency of NTIRE2024 Light Field Super Resolution Challenge and the seventh place in the Track 1 Fidelity.

In our future work, we will adopt model compression techniques, such as knowledge distillation, pruning, and model quantization, to further lighten our model and enhance its effectiveness.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China under Grant No.61901117, No.62301161, A21EKYN00884B03, in part by the Natural Science Foundation of Fujian Province under Grant No.2023J01083.

References

[1] Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy. Using plane+ parallax for calibrating dense camera arrays. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 1

[2] Yingqian Wang, Jungang Yang, Yulan Guo, Chao Xiao, and Wei An. Selective light field refocusing for camera arrays using bokeh rendering and superresolution. *IEEE Signal Processing Letters*, 26(1):204–208, 2018. 1

[3] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4748–4757, 2018. 1

[4] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818, 2022.

[5] Wentao Chao, Xuechun Wang, Yingqian Wang, Guanghui Wang, and Fuqing Duan. Learning sub-pixel disparity distribution for light field depth estimation. *IEEE Transactions on Computational Imaging*, 9:1126–1138, 2023. 1

[6] Ryan S Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr, and Paul Debevec. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 1

[7] Jingyi Yu. A light-field journey to virtual reality. *IEEE MultiMedia*, 24(2):104–112, 2017. 1

[8] Gaochang Wu, Yebin Liu, Lu Fang, and Tianyou Chai. Revisiting light field rendering with deep anti-aliasing neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5430–5444, 2021. 1

[9] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

[10] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *European Conference on Computer Vision*, pages 612–629. Springer, 2022.

[11] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19819–19829, 2022. 1

[12] Tom E Bishop and Paolo Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):972–986, 2011. 1, 3

[13] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):606–619, 2013. 3

[14] Reuben A Farrugia, Christian Galea, and Christine Guillemot. Super resolution of light field images using linear subspace projection of patch-volumes. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1058–1071, 2017. 3

[15] Mattia Rossi and Pascal Frossard. Graph-based light field super-resolution. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2017. 3

[16] Martin Alain and Aljosa Smolic. Light field denoising by sparse 5d transform domain collaborative filtering. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2017. 1, 3

[17] Nan Meng, Hayden K-H So, Xing Sun, and Edmund Y Lam. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):873–886, 2019. 1, 3

[18] 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*, pages 24–32, 2015. 1, 3

[19] 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, 3

[20] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11046–11055, 2019. 3, 5, 7, 8

[21] Zhen Cheng, Zhiwei Xiong, and Dong Liu. Light field super-resolution by jointly exploiting internal and external similarities. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2604–2616, 2019. 3

[22] 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, 2020. 3, 5, 6

[23] 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. 1, 2, 3, 7, 8

[24] 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, 7, 8

[25] 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: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 290–308. Springer, 2020. 3, 5, 7, 8

[26] 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, 3

[27] Gaosheng Liu, Huanjing Yue, Jiamin Wu, and Jingyu Yang. Intra-inter view interaction network for light field image super-resolution. *IEEE Transactions on Multimedia*, 25:256–266, 2021. 3

[28] 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, 45(1):425–443, 2022. 2, 3

[29] 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. 2, 3

[30] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12376–12386, 2023. 2, 3

[31] Kai Jin, Angulia Yang, Zeqiang Wei, Sha Guo, Mingzhi Gao, and Xiuzhuang Zhou. Distgepit: Enhanced disparity learning for light field image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1373–1383, 2023. 2, 3

[32] Yan Yuan, Ziqi Cao, and Lijuan Su. Light-field image super-resolution using a combined deep cnn based on epi. *IEEE Signal Processing Letters*, 25(9):1359–1363, 2018. 3, 7

[33] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10010–10019, 2021. 3

[34] Zeyu Xiao, Ruisheng Gao, Yutong Liu, Yueyi Zhang, and Zhiwei Xiong. Toward real-world light field super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3407–3417, 2023. 3

[35] Yingqian Wang, Zhengyu Liang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Real-world light field image super-resolution via degradation modulation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 3

[36] Zijian Wang and Yao Lu. Multi-granularity aggregation transformer for light field image super-resolution. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 261–265. IEEE, 2022. 3

[37] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 4

[38] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022.

[39] Xu Ma, Xiyang Dai, Jianwei Yang, Bin Xiao, Yinpeng Chen, Yun Fu, and Lu Yuan. Efficient modulation for vision networks. In *The Twelfth International Conference on Learning Representations*, 2024. 4

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

[41] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, pages 19–34. Springer, 2017. 6

[42] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, volume 13, pages 225–226, 2013. 6

[43] 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

[44] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7):3, 2008. 6

[45] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 7, 8

[46] 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*, pages 136–144, 2017. 7, 8

[47] 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. 7, 8

[48] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2260–2269, 2020. 7, 8

[49] Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte, Yulan Guo, et al. Ntire 2024 challenge on light field image super-resolution: Methods and results. In *CVPRW*, 2024. 7, 8