This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LFTramba: Comprehensive Information Learning for Light Field Image Super-Resolution via A Hybrid Transformer-Mamba Framework

Haosong Liu[1*] Xiancheng Zhu[1*] Huanqiang Zeng[1,2†]
Jianqing Zhu[1] Yifan Shi[1] Jing Chen[1] Junhui Hou[3]
[1] Huaqiao University [2] Xiamen University of Technology [3] City University of Hong Kong
{hsliu, xianchengzhu}@stu.hqu.edu.cn
{zeng0043, jqzhu, shiyifan, chenjing8005}@hqu.edu.cn jh.hou@cityu.edu.hk

Abstract

*Comprehensive exploration of information in complex 4D Light Fields (LF) is crucial for advancing Light Field image Super-Resolution (LFSR). Existing state-of-the-art methods based on CNN or Transformer typically decompose the 4D LF into 2D subspaces, limiting feature learning to specific aspects such as disparity cues or spatial-angular correlations. Therefore, they exhibit deficiencies in comprehensive subspace information integration and insufficient cross-domain feature interaction. Moreover, the potential of hybrid architectures that leverage multiple computational paradigms remains largely unexplored in LFSR. To overcome these limitations, we propose LFTramba, a novel hybrid network architecture that facilitates collaborative learning across multiple subspaces through a Transformer-Mamba framework. Specifically, the Spatial-Angular Mamba Block (SAMB), built on an efficient visual state space representation, alternates between capturing complementary spatial-angular features in the spatial and angular domains. Concurrently, the Epipolar Plane Transformer Block (EPTB) extracts disparity features alternately from horizontal and vertical Epipolar Plane Image (EPI) domains. Extensive experiments on real and synthetic datasets confirm that LFTramba surpasses existing LFSR methods while maintaining high computational efficiency. Notably, it achieved **3rd in Track 1** and **5th in Track 2** of the NTIRE 2025 Light Field Image Super-Resolution Challenge, highlighting its superior performance.*

1. Introduction

LF photography [2] captures not only light intensity but also directional information, enabling the reconstruction of comprehensive 3D scene representations. This capability supports a wide range of applications, including refocusing [14, 34, 53], depth estimation [15, 30, 52] and 3D reconstruction [47]. The advent of portable LF cameras (*e.g.*, Lytro and Raytrix) has significantly broadened the scope of LF imaging applications. However, enhancing angular resolution in such devices often comes at the cost of spatial resolution, thereby constraining their practical utility. As a result, the development of efficient LFSR algorithms has emerged as a key research focus.

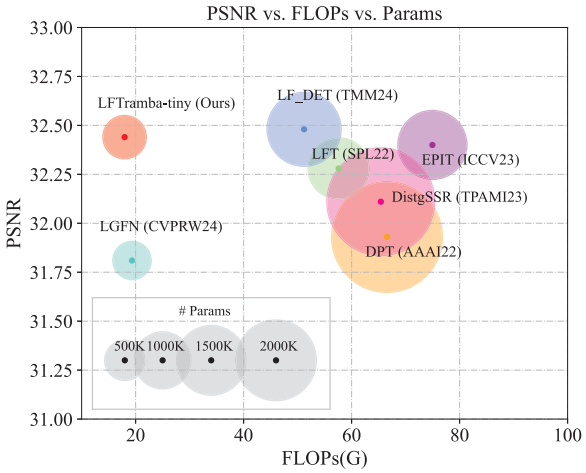


Figure 1. Comparisons of the PSNR, FLOPs and parameters of different LFSR methods.

This study focuses on enhancing the spatial resolution of LF images. Existing approaches tackle this challenge through various technical paradigms. Traditional methods [3, 6, 28, 41] formulate LFSR as an optimization problem, leveraging disparity estimation for multi-view image warping. However, these methods struggle to deliver robust performance when applied to low-resolution LF data and complex imaging scenarios involving occlusions and non-

*Equal contribution.

†Corresponding author. This work was supported by the Key Science and Technology Project of Fujian Province under Grant 2024HZ022007.

Lambertian reflections [25], ultimately limiting their reconstruction quality. In contrast, deep learning-based methods have become the dominant paradigm by exploiting spatial-angular features and disparity cues, with Transformer-based architectures [7, 13, 20, 21, 33] surpassing CNN-based approaches [16, 23, 26, 32, 35, 36, 38, 44, 49, 50] and achieving state-of-the-art super-resolution performance.

Nevertheless, existing methods face critical challenges in the holistic exploration and comprehensive learning of LF information. First, most approaches are constrained to single-domain learning, either capturing spatial-angular features or extracting disparity features from the EPI domain. For instance, EPIT [21] is specialized in EPI domain feature extraction, whereas LFT [20] and LF-DET [7] primarily focus on spatial-angular analysis. This limitation arises from the quadratic computational complexity inherent to Transformer architectures, which restricts simultaneous modeling across all subspaces. Second, the synergistic benefits of hybrid architectural paradigms in LFSR remain largely underexplored. Notably, DistgEPIT [17], the champion of the inaugural NTIRE LFSR challenge [37], empirically validated the effectiveness of hybrid designs, demonstrating their superiority in enhancing LFSR performance.

To address these limitations, we propose LFTramba, a novel hybrid architecture that combines the linear computational efficiency of Mamba with the global modeling capability of Transformer to enable comprehensive subspace feature learning. Due to the considerably longer unfolding sequence lengths in the spatial-angular domain, we employ Mamba blocks with selective state space for efficient linear complexity processing. In contrast, the compact sequences in the EPI domain, which exhibit critical long-range dependencies, are better suited for self-attention driven Transformer modules. Specifically, we first employ the Light Field Mamba Block (LFMB) to construct the Spatial-Angular Mamba Block (SAMB), enabling complementary information extraction in both spatial and angular domains. The LFMB integrates the Effective Visual State Space (EVSS) module with the channel attention mechanism, enhancing feature discrimination through spatial and channel path information interactions. Subsequently, the Epipolar Plane Transformer Block (EPTB) precisely captures disparity-sensitive features in the EPI domain, enhancing fine-structure recovery. Notably, the output features from SAMB serve as an intermediate representation that guides and supplements the EPTB, ensuring that the disparity-aware modeling benefits from enriched spatial-angular context. This interconnection facilitates an information flow wherein SAMB provides robust low-level spatial-angular priors, while EPTB refines them through domain-specific epipolar reasoning. Finally, a cross-domain fusion module hierarchically integrates multi-subspace features, promoting complementary feature aggregation across domains. As shown in Fig. 1, our method achieves comparable reconstruction accuracy in the low-parameter LFTramba-tiny, while the higher-parameter LFTramba further optimizes non-local spatial-angular correlation modeling and demonstrates enhanced robustness in large disparity scenarios.

In summary, the contributions of this paper are as follows: (1) We introduce LFTramba, a hybrid architecture combining Transformer and Mamba, to jointly model spatial, angular and disparity information across all LF subspaces. (2) We design the efficient SAMB for learning complementary spatial-angular features and propose the EPTB for extracting intrinsic disparity information, highlighting the beneficial role of spatial-angular features in disparity recovery. (3) Extensive experiments on public datasets confirm that LFTramba surpasses state-of-the-art methods in both reconstruction accuracy and computational efficiency. Moreover, it achieved 3rd place in Track 1 and 5th place in Track 2 of the NTIRE 2025 LFSR Challenge.

2. Related Work

LFSR aims to reconstruct high-resolution 4D LF structures from low-resolution sampling, with the primary challenge lying in simultaneously enhancing spatial detail fidelity and cross-view consistency. Existing research methodologies are broadly categorized into traditional methods and deep learning-based methods.

2.1. Traditional Methods

Traditional LFSR methods can be categorized into depth-dependent and depth-independent approaches. Depth-dependent approaches primarily focus on scene disparity/depth estimation and optimization frameworks: Bishop *et al.* [3] established a variational Bayesian inference framework that enhances view consistency through depth map recovery with Lambertian and texture-preserving priors. Wanner *et al.* [41] constructed local disparity models via EPI domain analysis, while Rossi *et al.* [28] optimized geometric reconstruction through graph regularization constraints. Depth-independent approaches employ data-driven strategies: Cho *et al.* [6] developed sparse representation dictionary learning for SAI super-resolution, Alain *et al.* [1] designed frequency-domain filtering to suppress ringing artifacts, and Boominathan *et al.* [4] proposed a hybrid imaging system integrating low-resolution LF cameras with high-resolution DSLRs. While these traditional methods demonstrate strong physical interpretability, they inherently suffer from accumulated depth estimation errors and limited generalization capabilities in complex scenarios.

2.2. Deep Learning-based Methods

Leveraging the powerful representational capacity of modern neural networks, deep learning-based LFSR approaches

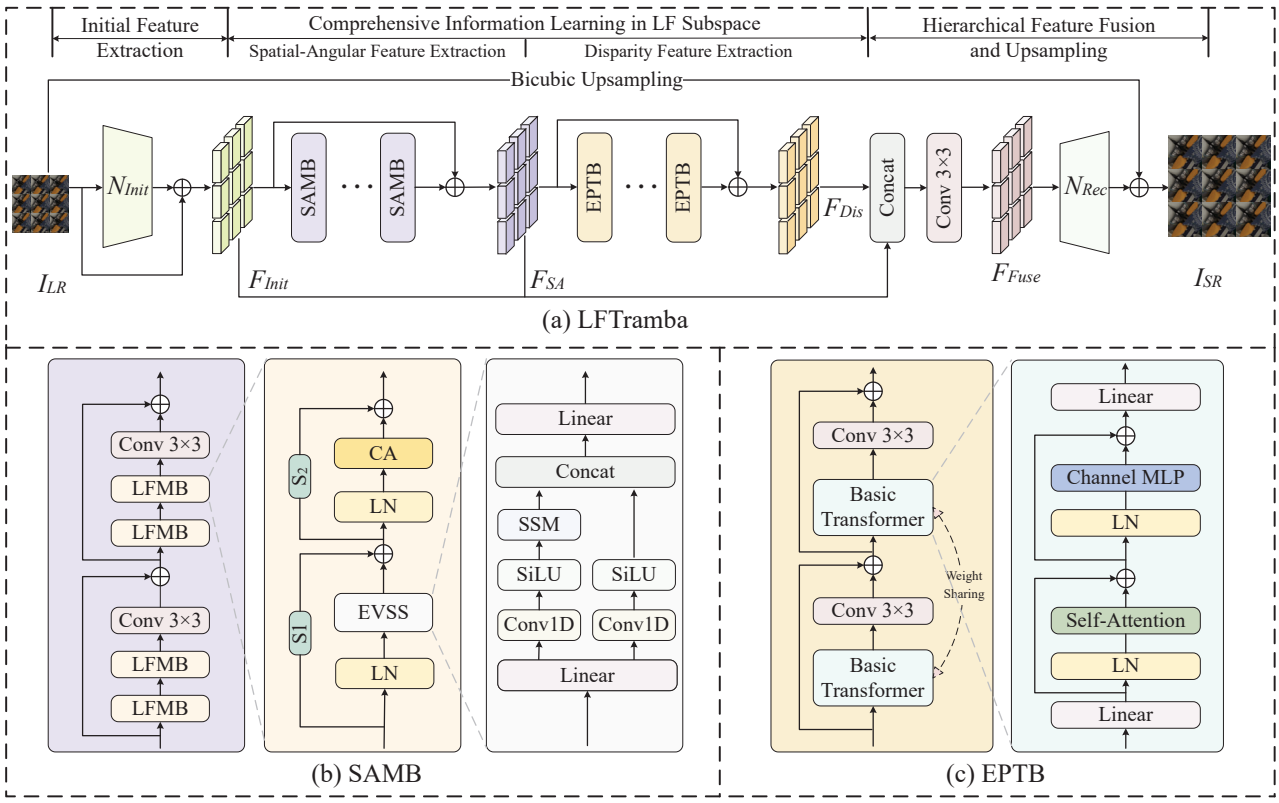


Figure 2. (a) The network architecture of the proposed LFTramba; (b) Illustration of the proposed Spatial-Angular Mamba Block (SAMB); (c) Illustration of the proposed Epipolar Plane Transformer Block (EPTB).

have achieved remarkable progress. CNN-based approaches were initially adapted to LFSR to learn local spatial-angular features for reconstruction enhancement: Yoon *et al.* [45] and S. Zhang *et al.* [49] exploited local spatial-angular features among adjacent views through basic CNNs. Subsequent advancements include Yeung *et al.*'s 4D convolution framework [44], Jin *et al.*'s multi-view fusion mechanism [16] and Wang *et al.*'s spatial-angular decoupled interaction networks [35, 38] for multidimensional feature modeling. Recent advancements, including Van Duong *et al.* [32], Yu *et al.* [46] and Zou *et al.* [55], have further improved feature extraction and reconstruction quality. Nevertheless, CNN-based methods remain fundamentally constrained by local receptive fields, struggling to effectively model cross-view long-range dependencies.

In recent years, Transformer-based architectures have been increasingly adopted in LFSR, leveraging the self-attention mechanism to model long-range dependencies across views and spatial dimensions. Liang *et al.* [20] introduced the attention mechanism for spatial-angular interactions, while Wang *et al.* [33] developed gradient-guided sequence learning to enhance detail recovery. Liang *et al.* [21] improved adaptability in large-disparity scenarios through EPI sequence modeling, and Cong *et al.* [7] optimized computational efficiency with subsampling convolutions and multi-scale angular processing. While these methods address CNN limitations in local perception, the quadratic complexity of self-attention hinders efficient feature learning across all subspaces.

Recent advancements have progressively validated the methodological merits of hybrid architectures. The representative DistgEPIT [17] achieved championship in the inaugural LFSR challenge by synergizing CNN for local spatial-angular feature extraction with Transformer for non-local disparity modeling in EPI domain. While hybrid architectures demonstrate breakthrough potential, their theoretical frameworks and implementation paradigms require deeper exploration. Notably, visual state space models [10, 11, 19, 24, 54] present a novel paradigm for processing 4D LF data through their synergistic advantages of linear complexity and global receptive fields. Our work pioneers the integration of visual state space models with Transformer architectures in LFSR. By innovatively devising a state space-Transformer hybrid framework, we achieve comprehensive LF information learning to significantly enhance reconstruction fidelity.

Table 1. Quantitative comparison (PSNR/SSIM) of different methods for $\times 4$ LFSR with 5×5 angular resolution. The best and second-best results are highlighted in red and blue, respectively.

Method	Params.(M)	FLOPs(G)	EPFL	HCInew	HCIold	INRIA	STFgantry	Average
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	-	-	25.26/0.8324	27.72/0.8517	32.58/0.9344	26.95/0.8867	26.09/0.8452	27.72/0.8701
EDSR [22]	38.89	1017	27.84/0.8854	29.60/0.8869	35.180/.9536	29.66/0.9257	28.70/0.9072	30.20/0.9118
RCAN [51]	15.36	408.5	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 [49]	8.64	39.70	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 [44]	1.77	128.44	28.27/0.9118	30.72/0.9145	36.70/0.9696	30.31/0.9467	30.15/0.9426	31.52/0.9370
LF-InterNet [35]	5.48	50.10	28.67/0.9162	30.98/0.9161	37.11/0.9716	30.64/0.9491	30.53/0.9409	31.58/0.9388
LF-ATO [16]	1.36	686.99	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-IINet [23]	4.88	57.42	29.11/0.9188	31.36/0.9208	37.62/0.9734	31.08/0.9515	31.21/0.9502	32.08/0.9429
DistgSSR [38]	3.58	65.41	28.99/0.9195	31.38/0.9217	37.56/0.9732	30.99/0.9519	31.65/0.9535	32.11/0.9440
HLFSR [32]	13.87	182.52	29.20/0.9222	31.57/0.9238	37.78/0.9742	31.24/0.9534	31.64/0.9537	32.29/0.9455
LFT [20]	1.16	57.60	29.25/0.9210	31.46/0.9218	37.63/0.9735	31.20/0.9524	31.86/0.9548	32.28/0.9447
DPT [33]	3.78	66.55	28.93/0.9170	31.19/0.9188	37.39/0.9721	30.96/0.9503	31.14/0.9488	31.93/0.9414
EPIT [21]	1.47	74.96	29.34/0.9197	31.51/0.9231	37.68/0.9737	31.37/0.9526	32.18/0.9571	32.40/0.9452
LF-DET [7]	1.69	51.20	29.47/0.9230	31.56/0.9235	37.84/0.9744	31.39/0.9534	32.14/0.9573	32.48/0.9463
DistgEPIT [17]	19.02	397.20	30.09/0.9224	31.61/0.9252	37.96/0.9742	32.35/0.9535	32.45/0.9589	32.90/0.9468
LFTramba-tiny	0.58	19.01	29.52/0.9206	31.48/0.9226	37.62/0.9735	31.60/0.9526	31.96/0.9563	32.44/0.9451
LFTramba	11.20	340.80	30.07/0.9284	31.90/0.9277	38.14/0.9756	32.14/0.9564	32.75/0.9609	33.00/0.9498
LFTramba*	11.20	340.80	30.21/0.9303	32.11/0.9297	38.33/0.9765	32.20/0.9576	33.02/0.9627	33.18/0.9514

Notes: * denotes the TTA technique and PSW test scheme [17].

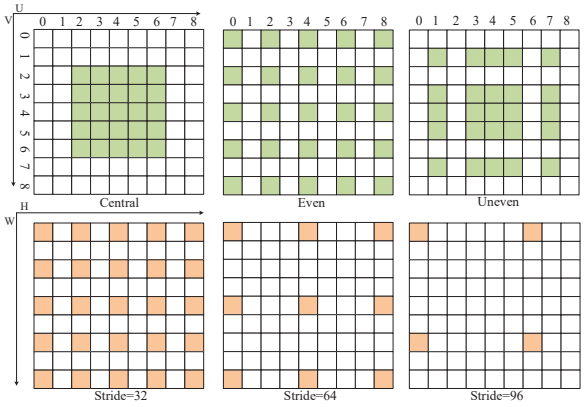


Figure 3. Illusatrtion of the Stride-Optimized Data Resampling (SODR), including central, even, and uneven sampling with strides of 32, 64 and 96, respectively.

3. Method

3.1. Network Architecture

As shown in Fig. 2(a), LFTramba consists of three main components: initial feature extraction, comprehensive information learning in LF subspace, hierarchical feature fusion and upsampling. The network takes a low-resolution LF image $I_{LR} \in \mathbb{R}^{U \times V \times H \times W}$ as input and reconstructs the high-resolution LF image $I_{HR} \in \mathbb{R}^{U \times V \times \alpha H \times \alpha W}$, where $U \times V$ is the angular resolution, $H \times W$ is the spatial resolution, and α is the scale factor.

Initial Feature Extraction: Similar to previous works [21, 38], an encoder $\mathcal{N}_{Init}$ composed of cascaded convolutional layers is employed to extract initial features $F_{Init} \in \mathbb{R}^{U \times V \times H \times W \times C}$ from the input I_{LR}:

$$F_{Init} = \mathcal{N}_{Init}(I_{LR}), \tag{1}$$

where C denotes the channel dimension.

Comprehensive Information Learning in LF Subspace: To comprehensively capture information from all LF subspaces, F_{Init} is first fed into a sequence of Spatial-Angular Mamba Blocks (SAMB). These blocks alternately extract spatial-angular features across the spatial and angular domains, producing enriched features $F_{SA} \in \mathbb{R}^{U \times V \times H \times W \times C}$ that incorporate spatial-angular correlations:

$$F_{SA} = SAMB^{i}(SAMB^{i-1}(\dots SAMB^{1}(F_{Init}))) + F_{Init}, \tag{2}$$

where $SAMB(\cdot)$ denotes the function of SAMB.

Next, F_{SA} is processed by a series of Epipolar Plane Transformer Blocks (EPTB) to capture disparity-sensitive features by interacting across the horizontal and vertical EPI domains. This step enhances the representation by embedding LF disparity information, yielding $F_{Dis} \in \mathbb{R}^{U \times V \times H \times W \times C}$:

$$F_{Dis} = EPTB^{j}(EPTB^{j-1}(\dots EPTB^{1}(F_{SA}))) + F_{SA}, \tag{3}$$

where $EPTB(\cdot)$ represents the function of EPTB.

Hierarchical Feature Fusion and Upsampling: The hierarchical fusion stage concatenates F_{Init}, F_{SA}, F_{Dis} along the channel dimension, followed by convolution operation $Conv$ to produce the fused feature $F_{Fuse} \in \mathbb{R}^{U\times V\times H\times W\times 3C}$:

$$F_{Fuse} = Conv(Concat(F_{Init}, F_{SA}, F_{Dis})). \quad (4)$$

Finally, the fused feature F_{Fuse} is fed into a decoder $\mathcal{N}_{Rec}$ composed of pixel-shuffle [29] and convolution layers to generate the high-resolution LF image I_{HR}:

$$I_{HR} = \mathcal{N}_{Rec}(F_{Fuse}). \quad (5)$$

3.2. Spatial-Angular Mamba Block

The structure of the SAMB is illustrated in Fig. 2(b). We adopt a simple yet effective spatial-angular separable modeling approach. Specifically, the input feature map $F \in \mathbb{R}^{U\times V\times H\times W\times C}$ is first reshaped into $F_{SAI} \in \mathbb{R}^{UV\times H\times W\times C}$. This reshaped feature is then processed by the Light Field Mamba Block (LFMB) followed by convolution to capture spatial contextual information, producing the spatially enhanced feature $F_S \in \mathbb{R}^{U\times V\times H\times W\times C}$:

$$F_S = Conv(LFMB^2(LFMB^1(F_{SAI}))) + F_{SAI}. \quad (6)$$

Next, F_S is reshaped into $F_{MacPI} \in \mathbb{R}^{HW\times U\times V\times C}$, and it undergoes a similar process through another sequence of LFMB and convolutional layers to extract angular complementary information. This operation yields the enhanced angular feature $F_A \in \mathbb{R}^{U\times V\times H\times W\times C}$:

$$F_A = Conv(LFMB^2(LFMB^1(F_{MacPI}))) + F_{MacPI}, \quad (7)$$

where $LFMB(\cdot)$ denotes the function of LFMB.

Inspired by [8, 43], the LFMB is designed based on the Efficient Visual State Space (EVSS) model, with the addition of a channel attention layer to enhance inter-channel feature interactions. Compared to prior state space models [9, 24, 54], the architecture of LFMB exhibits several structural modifications. Specifically, we replace the causal convolution with a standard convolution to eliminate the unidirectional constraint. Additionally, a symmetric branch, consisting of an extra convolutional layer and SiLU activation function, is introduced to mitigate potential information loss caused by the sequential modeling of SSMs. The outputs from these two branches are concatenated and projected through a linear layer, facilitating the fusion of sequential and spatial information and leveraging their complementary advantages. Given the input feature F_{in}, the output F_{out} of EVSS is computed as follows:

$$\begin{aligned} F_1 &= \text{SSM}(\sigma(\text{Conv}(\text{Linear}(C, \frac{C}{2})(F_{in})))), \\ F_2 &= \sigma(\text{Conv}(\text{Linear}(C, \frac{C}{2})(F_{in}))), \\ F_{out} &= \text{Linear}(C, C)(\text{Concat}(F_1, F_2)), \end{aligned} \quad (8)$$

where SSM refers to the Selective State Space Model described in [9], σ represents the SiLU activation function, and Linear denotes the linear projection layer.

3.3. Epipolar Plane Transformer Block

The detailed structure of the EPTB is illustrated in Fig. 2(c). Its design is inspired by the Non-Local Cascading Block proposed in [21], with appropriate modifications to enhance global disparity perception across the EPI domain. To strengthen the representation of local features, EPTB incorporates single-layer spatial convolution. This is because multiple convolutional layers are deliberately avoided in order to preserve the inherent parallax structure of the light field. Specifically, each EPTB alternately extracts disparity information from the horizontal and vertical EPI domains using Basic Transformer. Notably, the two Basic Transformer units within each module share weights, a design choice that facilitates more effective capture of the spatial-angular correlations.

3.4. Stride-Optimized Data Resampling

Current LF datasets commonly suffer from skewed disparity distributions characterized by scarcity of large-disparity samples and imbalanced data representation, which severely constrains model generalization in wide-baseline scenarios. Although previous studies [5, 17] have employed central, even, and uneven sampling strategies for data augmentation to increase large-disparity samples, these approaches often lead to a significant increase in training time. To tackle this issue, we introduce a Stride-Optimized Data Resampling (SODR) approach, which incorporates a progressive multi-stride sampling technique. As illustrated in Fig. 3, central, even and uneven resampling operate with strides of 32, 64, and 96 respectively. This configuration constructs a sampling distribution balancing large-disparity coverage and scene texture diversity. Compared with the augmented sampling in [5, 17], our approach enhances robustness to disparity variations while reducing training time and GPU memory consumption.

4. Experiments

In this section, we first provide a detailed description of the dataset and experimental settings, followed by a comprehensive performance comparison between our method and existing state-of-the-art methods. Subsequently, extensive ablation studies are conducted to validate the effectiveness and robustness of LFTramba.

4.1. Datasets and Implementation Details

Datasets: In this study, we follow the experimental setup of existing works [20, 21, 38] and select five widely used benchmark datasets for LFSR: EPFL [27], HCIold [42], HCInew [12], INRIA [18] and STFgantry [31]. As shown

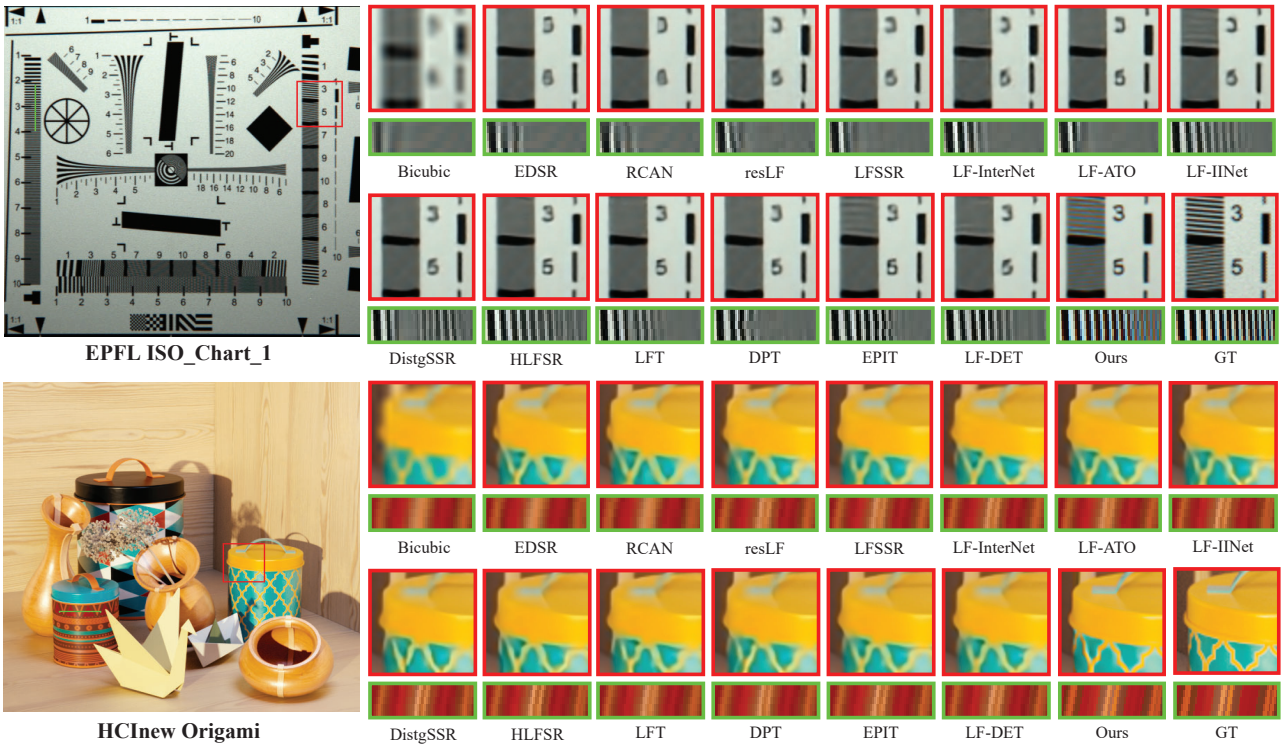


Figure 4. Comparison of visual results for the × 4 LFSR with state-of-the-art methods. As shown in the figure, the enlarged spatial patches of the center SAIs are highlighted with red boxes, and the enlarged EPI slices are highlighted with green boxes.

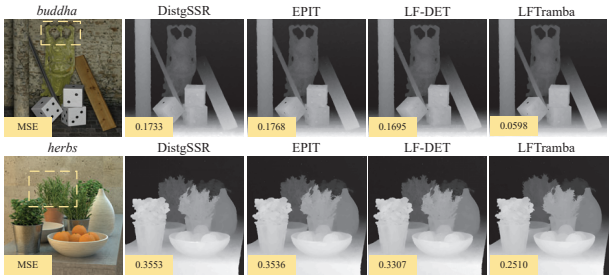


Figure 5. Depth estimation results obtained by SPO [48] using × 4 SR images from different LFSR methods. We report the mean square error multiplied by 100 (MSE × 100), where lower values are better.

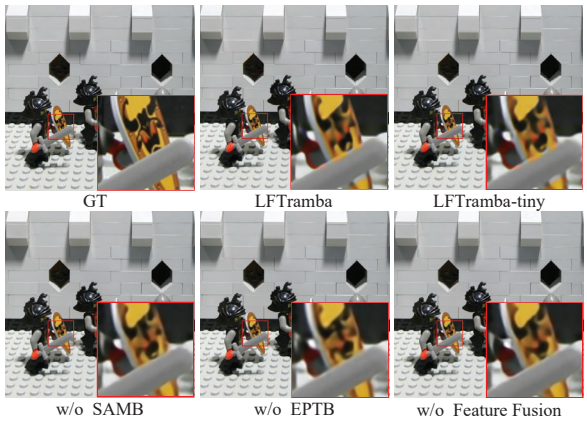


Figure 6. Visualization of the impact of LFTramba-tiny without SAMB, EPTB or Feature Fusion on the results of × 4 LFSR.

in Fig. 3, for training, we employ the SODR strategy to extract 5×5 SAIs, which are then cropped into 64×64 high-resolution image patches. Corresponding 16×16 low-resolution image patches are generated using bicubic downsampling. During testing, we extract only the central 5×5 SAIs.

Implementation Details: To improve the model's generalization, we apply data augmentation techniques, including random horizontal and vertical flips and 90° random rotations, during training. Both LFTramba and LFTramba-tiny are trained using the L1 loss and optimized with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The initial learning rate is 2×10^{-4}, halved every 15 epochs. LFTramba is configured with 2 SAMBs, 10 EPTBs, a channel size of 128 and a batch size of 16, trained for 75 epochs. And LFTramba-tiny is configured with 4 SAMBs, 3 EPTBs, a channel size of 32 and a batch size of 2, trained for 90

Table 2. Ablation results (PSNR/SSIM) on the Stride-Optimized Data Resampling (SODR) for $\times$ 4 LFSR.

SODR	Time Cost/Epoch	EPFL	HCInew	HCIold	INRIA	STFgantry	Average
w/o	0.93h	29.49/0.9201	31.47/0.9223	37.60/0.9732	31.48/0.9522	31.81/0.9545	32.37/0.9444
w	1.43h	29.52/0.9206	31.48/0.9226	37.62/0.9735	31.60/0.9526	31.96/0.9563	32.44/0.9451

Table 3. Ablation results (PSNR/SSIM) on the components and structure of LFTramba for $\times$ 4 LFSR. The bolded row represents the baseline used by LFTramba.

Cascade Order	#Params	FLOPs	SAMB	EPTB	Fusion	EPFL	HCInew	HCIold	INRIA	STFgantry	Average
SAMB+ EPTB	**0.58**	**17.96**	✓	✓	✓	**29.52/0.9206**	**31.48/0.9226**	**37.62/0.9735**	**31.60/0.9526**	**31.96/0.9563**	**32.44/0.9451**
EPTB+SAMB	0.58	17.96	✓	✓	✓	29.45/0.9202	31.42/0.9217	37.69/0.9736	31.37/0.9520	31.78/0.9540	32.40/0.9443
SAMB+ EPTB	0.55	17.12	✓	✓		29.41/0.9198	31.49/0.9223	37.55/0.9730	31.48/0.9522	31.80/0.9546	32.35/0.9444
-	0.61	16.71	✓		✓	29.39/0.9205	31.47/0.9222	37.62/0.9737	31.32/0.9521	31.64/0.9537	32.29/0.9444
-	0.55	25.10		✓	✓	29.01/0.9162	31.32/0.9208	37.48/0.9727	31.05/0.9507	31.99/0.9554	32.17/0.9432

epochs. All experiments are conducted in PyTorch on a single NVIDIA RTX A6000 GPU.

During inference, we apply the Position Sensitive Window (PSW) operation [17] and Test Time Augmentation (TTA) to further enhance reconstruction quality. For performance evaluation, we use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [40] as metrics, computed on the Y channel. To obtain overall performance, we first average the scores across all SAIs in each scene, then compute the mean across all scenes in the dataset.

4.2. Comparison with State-of-the-art methods

We compare LFTramba with several state-of-the-art methods, including single-image SR methods [22, 51], CNN-based LFSR methods [16, 23, 32, 35, 38, 44, 49], Transformer-based LFSR methods [7, 20, 21, 33] and the hybrid CNN-Transformer method [17].

Quantitative Results: Table 1 presents a quantitative comparison of LFSR methods, where our LFTramba achieves state-of-the-art performance, exceeding DistgEPIT [17] by 0.1 dB in PSNR. Furthermore, the incorporation of TTA and PSW [17] further improves performance, achieving a PSNR of 33.18 dB and an SSIM of 0.9514. Notably, the lightweight LFTramba-tiny model maintains competitive accuracy with significantly fewer parameters, outperforming EPIT [21] by 0.04 dB and LFT [20] by 0.16 dB. Fig. 1 illustrates the trade-off between performance and model complexity, demonstrating the superior efficiency of our approach.

Qualitative Results: Fig. 4 presents qualitative results for $\times$4 LFSR across different methods. The magnified regions highlight our approach's ability to restore fine details and sharpen edges. For instance, in the ISO_Chart_1 scene, our method accurately reconstructs the number '3' and '5', preserving their intricate structures more clearly than other methods while effectively mitigating severe blurring. In the $Origami$ scene, it restores the intersecting straight lines on the jar's surface with greater precision, significantly reducing blurring and artifacts compared to competing approaches. Additionally, LFTramba demonstrates superior performance in EPI visualization, generating sharper, clearer lines with fewer artifacts.

Angular Consistency: Additionally, we assess the angular consistency of reconstructed LF images using the SPO [48] depth estimation algorithm. As shown in Fig. 5, depth maps produced by other LFSR methods exhibit noticeable noise and high MSE$\times$100 errors, indicating their limited ability to effectively integrate spatial-angular and disparity information. In contrast, our approach significantly improves depth estimation accuracy, producing clearer structures with the lowest MSE$\times$100 error. This further validates that the high-resolution LF images generated by LFTramba have superior angular consistency.

4.3. Ablation Study

In this subsection, we further analyze the effectiveness of the strategies and architecture proposed in our work, including the SODR strategy and the core components and structure of LFTramba.

Stride-Optimized Data Resampling. As shown in Table 2, the introduction of SODR strategy significantly enhances the model performance on large disparity dataset (*e.g.*, STFgantry [31]), yielding a notable improvement of 0.15 dB. This strategy also demonstrates consistent performance improvements across other datasets, highlighting its robustness in adapting the model to both large disparity scenarios and small-to-medium disparity scenes. Importantly, the additional training overhead introduced by this method is minimal, increasing the training time per epoch by only approximately 50%, while significantly improving recon-

Table 4. Our team (Bold Name) secured 3rd place in Track 1 and 5th place in Track 2 on the NTIRE-2025 test dataset leaderboard [39].

	Rank	Team	Test Set			#Params	FLOPs
			Average	Lytro	Synthetic		
Track 1	1	OpenMeow	31.1638/.9366	31.4330/.9531	30.8946/.9201	9.10M	445.08G
	2	BITSMBU	31.0915/.9354	31.4713/.9522	30.7117/.9186	3.91M	115.75G
	3	**SmartVIPLab**	30.8576/.9336	31.0748/.9503	30.6403/.9170	11.20M	340.80G
Track 2	1	BITSMBU	30.3889/.9289	30.5700/.9454	30.2078/.9125	0.54M	17.03G
	2	OpenMeow	30.3440/.9280	30.5043/.9445	30.1837/.9116	0.45M	19.33G
	3	LFSR-DASE	30.2328/.9270	30.4136/.9434	30.0520/.9107	0.98M	19.87G
	4	CBNU-MIP&VC-Lab	30.1314/.9255	30.2604/.9419	30.0023/.9092	0.66M	19.88G
	5	**SmartVIPLab**	30.1310/.9258	30.2887/.9424	29.9733/.9092	0.58M	19.01G
Baselines	-	DistgEPIT [17]	30.66/.9314	30.82/.9475	30.51/.9152	20.34M	566.48G
	-	EPIT [21]	29.87/.9259	29.72/.9420	30.03/.9097	1.47M	76.39G
	-	DistgSSR [38]	29.64/.9244	29.39/.9403	29.88/.9084	3.58M	65.27G
	-	Bicubic	25.79/.8378	25.11/.8404	26.46/.8352	-	-

struction quality and yielding better results across various datasets.

Cascade Order of SAMB and EPTB modules. SAMB and EPTB are connected in a sequential manner. To investigate the impact of different ordering strategies, we design two configurations: prioritizing spatial-angular feature extraction (*i.e.*, SAMB+EPTB) and prioritizing disparity learning (*i.e.*, EPTB+SAMB). As shown in Table 3, the SAMB+EPTB structure outperforms EPTB+SAMB, indicating that extracting spatial-angular features first facilitates the recovery of intrinsic disparity information in LF. However, excessive emphasis on spatial-angular information may distort disparity structures to some extent, necessitating a balanced integration of both to optimize model performance.

LFTramba w/o SAMB, EPTB or Feature Fusion. To evaluate the contribution of each component in LFTramba, we remove SAMB, EPTB and Feature Fusion individually and retrain the network. To ensure a fair comparison, we appropriately deepen the network to maintain a similar parameters. As shown in Table 3, removing any component leads to a significant drop in PSNR for the $\times$ 4 LFSR. Specifically, excluding SAMB results in a 0.27 dB decline, while removing EPTB causes a degradation of 0.15 dB. Notably, as shown in the fourth row of Table 3, the SAMB-based network outperforms LFT [20] in reconstruction performance while significantly reducing both parameters and FLOPS. This demonstrates that LFMB provides strong representational power without compromising efficiency, underscoring its practical applicability. Similarly, omitting Feature Fusion leads to a 0.09 dB decrease. This is because SAMB effectively selects and integrates critical spatial-angular features, while EPTB enhances disparity representation in the EPI domain, reinforcing spatial-angular correlations. Feature Fusion further exploits hierarchical features through simple concatenation and convolution. Fig. 6 provides visual comparisons, highlighting the synergistic effect of SAMB and EPTB, and demonstrating their crucial role in achieving comprehensive information learning in LF subspaces, ultimately driving performance improvements in LFSR.

4.4. NTIRE 2025 LFSR Challenge Results

The NTIRE 2025 LFSR challenge [39] used a test set consisting of 16 synthetic and 16 real-world LFs. For **Track1: Classic** and **Track 2: Efficiency**, participants were strictly prohibited from using external models or data, including pre-trained backbones and optical flow networks. To obtain the final results, we applied an average ensemble of LFTramba, BigEPIT [5], DistgEPIT [17] and RR-HLFSR [32], leveraging their structural diversity. As presented in Table 4, our team achieved **3rd place in Track 1** and **5th place in Track 2** on the LFSR test dataset.

5. Conclusion

In this paper, we propose LFTramba, a hybrid network that integrates Transformer and Mamba architectures to enhance LFSR. By jointly modeling spatial, angular, and disparity information across all LF subspaces, our method overcomes the limitations of existing CNN and Transformer-based approaches. The SAMB and EPTB improve information learning by addressing challenges in integration and cross-domain feature interaction. Extensive experiments on real and synthetic datasets validate its effectiveness, demonstrating a superior balance between reconstruction accuracy and model complexity. Additionally, LFTramba ranked 3rd in Track 1 and 5th in Track 2 of the NTIRE 2025 LFSR Challenge. Future work will focus on refining feature fusion strategies, enhancing computational efficiency, and expanding model adaptability.

References

[1] Martin Alain and Aljosa Smolic. Light field super-resolution via lfbm5d sparse coding. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 2501–2505. IEEE, 2018. 2

[2] Tom Bishop and Paolo Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:972–986, 2012. 1

[3] Tom E Bishop, Sara Zanetti, and Paolo Favaro. Light field superresolution. In *2009 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2009. 1, 2

[4] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2014. 2

[5] Wentao Chao, Yiming Kan, Xuechun Wang, Fuqing Duan, and Guanghui Wang. Bigepit: Scaling epit for light field image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6187–6197, 2024. 5, 8

[6] Donghyeon Cho, Minhaeng Lee, Sunyeong Kim, and Yu-Wing Tai. Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3280–3287, 2013. 1, 2

[7] Ruixuan Cong, Hao Sheng, Da Yang, Zhenglong Cui, and Rongshan Chen. Exploiting spatial and angular correlations with deep efficient transformers for light field image super-resolution. *IEEE Transactions on Multimedia*, 26:1421–1435, 2023. 2, 3, 4, 7

[8] Ruisheng Gao, Zeyu Xiao, and Zhiwei Xiong. Mamba-based light field super-resolution with efficient subspace scanning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 531–547, 2024. 5

[9] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023. 5

[10] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. *arXiv preprint arXiv:2411.15269*, 2024. 3

[11] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 222–241. Springer, 2024. 3

[12] 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, 2017. 5

[13] Zeke Zexi Hu, Xiaoming Chen, Vera Yuk Ying Chung, and Yiran Shen. Beyond subspace isolation: Many-to-many transformer for light field image super-resolution. *arXiv:2401.00740*, 2024. 2

[14] Sakila S Jayaweera, Chamira US Edussooriya, Chamith Wijenayake, Panajotis Agathoklis, and Len T Bruton. Multi-volumetric refocusing of light fields. *IEEE Signal Processing Letters*, 28:31–35, 2020. 1

[15] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. Depth from a light field image with learning-based matching costs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):297–310, 2019. 1

[16] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2260–2269, 2020. 2, 3, 4, 7

[17] 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, 4, 5, 7, 8

[18] 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. 5

[19] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 237–255. Springer, 2024. 3

[20] 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, 4, 5, 7, 8

[21] 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 International Conference on Computer Vision (ICCV)*, pages 12376–12386, 2023. 2, 3, 4, 5, 7, 8

[22] 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. 4, 7

[23] 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. 2, 4, 7

[24] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *NeurIPS*, 2024. 3, 5

[25] 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. 2

[26] Yu Mo, Yingqian Wang, Chao Xiao, Jungang Yang, and Wei An. Dense dual-attention network for light field image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4431–4443, 2022. 2

[27] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *Proceedings of the 8th International Conference on Quality of Multimedia Experience*, pages 1–2, 2016. 5

[28] 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, 2

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[30] Yu-Ju Tsai, Yu-Lun Liu, Ming Ouhyoung, and Yung-Yu Chuang. Attention-based view selection networks for light-field disparity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12095–12103, 2020. 1

[31] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Comput. Graph. Lab., Stanford Univ.*, 6(7), 2008. 5, 7

[32] Vinh Van Duong, Thuc Nguyen Huu, Jonghoon Yim, and Byeungwoo Jeon. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Transactions on Computational Imaging*, 9:350–366, 2023. 2, 3, 4, 7, 8

[33] 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*, pages 1–9, 2022. 2, 3, 4, 7

[34] 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 Process. Lett.*, 26(1):204–208, 2019. 1

[35] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 290–308. Cham, Switzerland: Springer, 2020. 2, 3, 4, 7

[36] 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. 2

[37] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, Kai Jin, Zeqiang Wei, Angulia Yang, Sha Guo, et al. Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1320–1335, 2023. 2

[38] 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, 2023. 2, 3, 4, 5, 7, 8

[39] Yingqian Wang, Zhengyu Liang, Fengyuan Zhang, Lvli Tian, Longguang Wang, Juncheng Li, Jungang Yang, Radu Timofte, and Yulan Guo. Ntire 2025 challenge on light field image super-resolution: Methods and results. In *CVPR Workshops*, 2025. 8

[40] 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. 7

[41] 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. 1, 2

[42] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, pages 225–226, 2013. 5

[43] Wang Xia et al. Lfmamba: Light field image super-resolution with state space model. *arXiv:2406.12463*, 2024. 5

[44] 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, 2019. 2, 3, 4, 7

[45] 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. 3

[46] Zhongxin Yu, Liang Chen, Zhiyun Zeng, Kunping Yang, Shaofei Luo, Shaorui Chen, and Cheng Zhong. Lgfn: Lightweight light field image super-resolution using local convolution modulation and global attention feature extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6712–6721, 2024. 3

[47] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6505–6514, 2021. 1

[48] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 6, 7

[49] 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, 4, 7

[50] 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

[51] 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. 4, 7

[52] Yuchen Zhang, Wenrui Dai, Mingxing Xu, Junni Zou, Xiaopeng Zhang, and Hongkai Xiong. Depth estimation from light field using graph-based structure-aware analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):4269–4283, 2020. 1

[53] Hao Zhu, Qi Zhang, and Qing Wang. 4d light field superpixel and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6384–6392, 2017. 1

[54] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv:2401.09417*, 2024. 3, 5

[55] Rong Zou, Wenjie Dai, Shenghe Bai, and Senlin Mu. Edge-preserving light-field image super-resolution via feature affine transformation network. *Optics and Lasers in Engineering*, 172:107839, 2024. 3