Occlusion-Embedded Hybrid Transformer for Light Field Super-Resolution

Zeyu Xiao[1], Zhuoyuan Li[2], and Wei Jia[3]*

[1]National University of Singapore
[2]University of Science and Technology of China
[3]Hefei University of Technology

Abstract

Transformer-based networks have set new benchmarks in light field super-resolution (SR), but adapting them to capture both global and local spatial-angular correlations efficiently remains challenging. Moreover, many methods fail to account for geometric details like occlusions, leading to performance drops. To tackle these issues, we introduce OHT. This hybrid network leverages occlusion maps through an occlusion-embedded mix layer. It combines the strengths of convolutional networks and Transformers via spatial-angular separable convolution (SASep-Conv) and angular self-attention (ASA). SASep-Conv offers a lightweight alternative to 3D convolution for capturing spatial-angular correlations, while the ASA mechanism applies 3D self-attention across the angular dimension. These designs allow OHT to capture global angular correlations effectively. Extensive experiments on multiple datasets demonstrate OHT's superior performance.

Introduction

The 4D light field, capturing both angular and spatial information, has become increasingly vital in computer vision (Wu et al. 2017). Commercial light field cameras utilize a micro-lens array in front of the sensor, leading to a trade-off between angular and spatial resolutions (Ng et al. 2005; Levin, Freeman, and Durand 2008; Sheng et al. 2022). This limited spatial resolution constrains practical applications, making light field super-resolution (SR) a critical research focus. Since the advent of light field cameras, SR has garnered significant attention (Raj et al. 2016; Xiao et al. 2023b, 2021).

With the rise of deep learning, convolutional neural network (CNN) based methods have shown promising performance for light field SR, surpassing classic non-learning-based methods with notable gains (Alain and Smolic 2018; Liang and Ramamoorthi 2015). To leverage complementary information from different views, recent CNNs employ various mechanisms: adjacent-view combination (Yoon et al. 2017), view-stack integration (Zhang, Lin, and Sheng 2019; Jin et al. 2020; Zhang, Chang, and Lin 2021), bidirectional recurrent fusion (Wang et al. 2018), spatial-angular disentanglement (Cheng, Liu, and Xiong 2022; Liang et al. 2022; Wang et al. 2022c; Yeung et al. 2018), 4D convolutions (Meng et al. 2019), and data augmentation strategy (Xiao et al. 2023b; Mi and Yang 2024). However, these methods still fail to fully explore the global and local spatial-angular correlations inherent in light fields.

Starting from natural language processing (Ashish 2017), Transformer architectures (Dosovitskiy et al. 2020) have been applied to vision tasks such as color image restoration (Liang et al. 2021) and light field SR (Cong et al. 2023; Wang et al. 2022a; Liang et al. 2022, 2023). Leveraging multi-head self-attention, these methods excel in capturing non-local similarities and long-range dependencies, surpassing CNN-based approaches. However, their application to light field SR remains sub-optimal due to quadratic complexities or disruption of angular dependencies. Light fields exhibit fixed structures, such as relative intensity correlations across views, which are often overlooked by existing transformer-based approaches. Directly applying these methods with accounting for such structures is appropriate. Moreover, many existing methods overlook crucial geometric information, such as occlusions in light fields, leading to significant errors and suboptimal results (as illustrated in Figure 1). This oversight can lead to sub-optimal performance, especially in scenarios where occlusions need accurate handling in light field SR tasks.

In this paper, we propose an Occlusion-Embedded Hybrid Transformer (OHT) that effectively integrates the local spatial-angular inductive bias of convolution and the long-range spatial-angular dependency modeling ability of Transformers. Unlike previous light field SR methods, OHT explicitly utilizes occlusion maps. Specifically, we obtain occlusion maps for each view by leveraging the photo-consistency constraint based on the coarse estimated disparity map using OACCNet (Wang et al. 2022b) from low-resolution (LR) observations (as shown in Figure 1. These occlusion maps represent the importance of pixels from different views, helping to mitigate the impact of occluded pixels. We incorporate this information into the SR process using the occlusion-embedded mix layer. In addition, OHT incorporates several key designs: (1) We introduce a spatial-angular separable convolution (SASep-Conv), which extracts spatial-angular features more effectively than previous methods like 2D convolutions (Jin et al. 2020), 3D convolutions (Tran, Berberich, and Simon 2022), 4D convolu-

*Corresponding Author

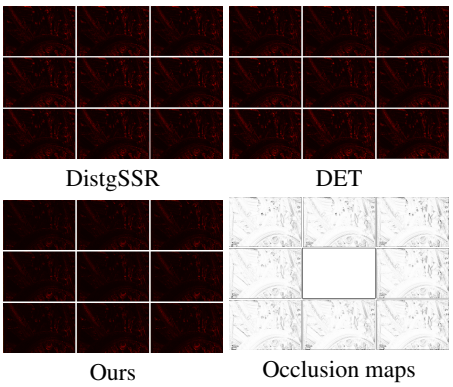

Figure 1: Examples of light field SR. We sequentially present images from angle coordinates (0,0), (0,2), (0,4), (2,0), (2,2), (2,4), (4,0), (4,2), and (4,4). From left to right: error maps of DistgSSR, error maps of DET, error maps of OHT-B (Ours), and the generated occlusion maps.

tions (Meng et al. 2019) and separable convolutions (Cheng, Liu, and Xiong 2022). (2) We process the extracted features with an angular self-attention (ASA) mechanism to leverage the enhanced local spatial-angular inductive bias. ASA mechanism performs global self-attention across the angular space, selectively aggregating information across different angular views. This approach enriches our model with robust capabilities to capture long-range angular correlations, mitigating the limitations of 2D angular self-attention and the quadratic complexity of spatial attention methods. Considering the importance of multi-view information from the angular dimension, this design ensures that the model effectively leverages this data for superior SR performance. (3) We introduce a convolutional feed-forward layer (CFFL) to further enhance aggregated features from informative angular regions. These designs address the aforementioned challenges of light field SR, effectively leveraging both local spatial-angular details and global angular correlations for superior performance and adaptability. As illustrated in Figure 1, OHT can generate results with fewer errors. Extensive experiments on benchmark datasets show that our OHT can achieve superior performance.

The contributions of this work are summarized as follows: (1) We present OHT, a hybrid spatial-angular transformer that effectively captures local spatial-angular features and long-range global angular correlations. (2) We introduce ASA, modeling long-range angular correlations along angular space instead of previous 2D spatial/channel dimensions. (3) We propose CFFL to drive the model to focus on more informative regions and SASep-Conv to extract meaningful spatial-angular features.

Related Work

Light field image super-resolution. Light field image SR is a long-standing, ill-posed problem. Traditional non-learning methods rely on geometric (Liang and Ramamoorthi 2015; Rossi and Frossard 2017) and mathematical (Alain and Smolic 2018) modeling of the 4D light field structure for super-resolution through projection and optimization techniques. Deep learning methods have recently become dominant due to their superior performance. Yoon *et al.* (Yoon et al. 2017) introduce LFCNN, the first light field SR network, adapting the SRCNN architecture (Dong et al. 2014) with multiple channels. Subsequent methods exploit across-view redundancy in the light field, either explicitly (Cheng, Xiong, and Liu 2019; Jin et al. 2020; Wang et al. 2018; Zhang, Lin, and Sheng 2019) or implicitly (Meng et al. 2019; Wang et al. 2020; Yeung et al. 2018; Yuan, Cao, and Su 2018; Wang et al. 2022c; Van Duong et al. 2023b,a; Liu, Yue, and Yang 2024; Xiao et al. 2023a; Xiao, Cheng, and Xiong 2023; Xiao, Shou, and Xiong 2024; Xiao and Xiong 2024; Xiao et al. 2024; Tang et al. 2024; Li et al. 2024c,a,b). Transformer-based methods have recently shown effectiveness in light field SR (Liang et al. 2022; Wang et al. 2022a; Liang et al. 2023; Cong et al. 2023). There is also a trend toward designing deeper and wider networks (Jin et al. 2023; Van Duong et al. 2023b) to improve light field SR accuracy. Recently, novel architectures such as diffusion-based (Gao et al. 2024; Chao et al. 2023) and Mamba-based (Gao, Xiao, and Xiong 2024; Lu et al. 2024) methods have demonstrated their effectiveness in the light field SR task.

Vision Transformers. Transformer (Ashish 2017), originally introduced as a parallel and purely attention-based alternative to recurrent neural networks (Chung et al. 2014) in natural language processing, has been successfully adapted for high-level vision tasks (Dosovitskiy et al. 2020). Recognizing its powerful representation abilities, recent works have also applied transformers to low-level tasks like natural image SR (Chen et al. 2023) and image denoising (Tian et al. 2024). One key challenge in these methods is the quadratic complexity of the self-attention (SA) mechanism. To address this, SwinIR (Liang et al. 2021) adapts the Swin Transformer (Liu et al. 2021) by replacing global attention with a more efficient shift-window-based attention. Similarly, Uformer (Wang et al. 2022d) uses attention over non-overlapping patches and adopts a U-Net architecture for increased efficiency. While Transformer-based methods have shown effectiveness in light field SR (Liang et al. 2022; Wang et al. 2022a; Liang et al. 2023; Cong et al. 2023), they still face challenges in capturing both global and local spatial-angular correlations efficiently and flexibly. Moreover, these methods often neglect geometric details, such as occlusions, unique to light fields. Our OHT addresses these challenges by introducing SA-SepConv and AngSA, which effectively extract angular-correlated features. OHT ensures improved performance for light field SR.

Occlusion-Embedded Hybrid Transformer

The proposed OHT aims to recover a clean and sharp high-resolution (HR) light field $\boldsymbol{I}^{sr} \in \mathbb{R}^{\alpha H \times \alpha W \times U \times V}$ given LR observation $\boldsymbol{I}^{lr} \in \mathbb{R}^{H \times W \times U \times V}$

$$\boldsymbol{I}^{sr} = \mathrm{OHT}(\boldsymbol{I}^{lr}), \tag{1}$$

where H and W are spatial dimensions, U and V are angular dimensions, and α represents the magnification factor (*i.e.*, 2 and 4 in our experiments).

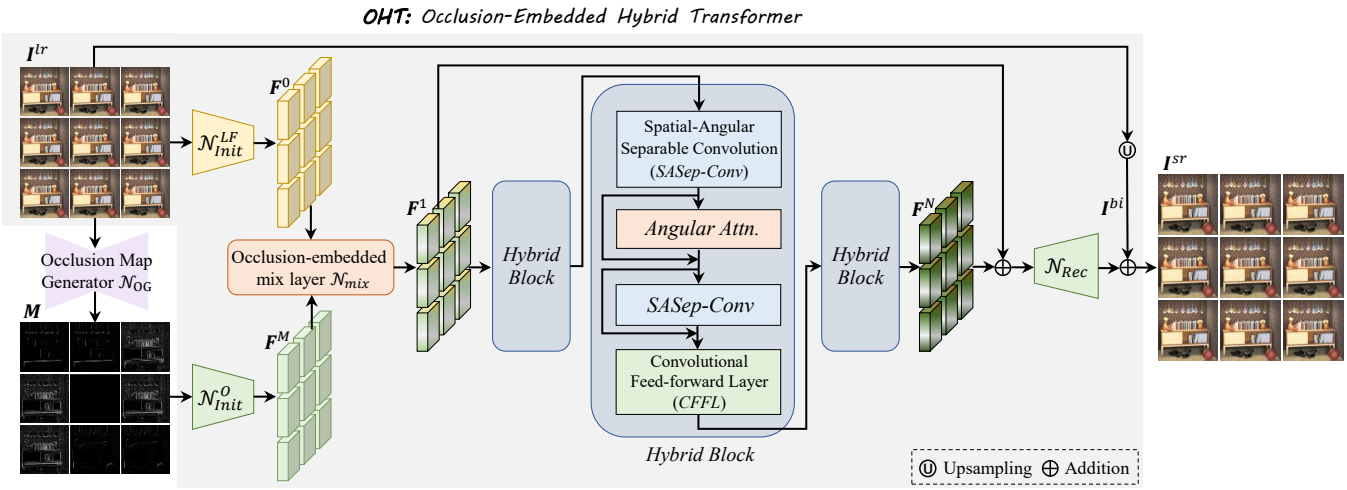

Figure 2: An overview of the proposed OHT. OHT is a hybrid network that leverages occlusion maps through an occlusion-embedded mix layer and combines the strengths of convolutional networks and Transformers via SASep-Conv and ASA.

The structure of OHT is shown in Figure 2. Given LR observation $\boldsymbol{I}^{lr}$, we first feed $\boldsymbol{I}^{lr}$ to the occlusion map generator $\mathcal{N}_{OG}(\cdot)$ and generate the occlusion maps $\boldsymbol{M}$. We then feed $\boldsymbol{I}^{lr}$ and $\boldsymbol{M}$ to initial feature extractors $\mathcal{N}^{LF}_{Init}(\cdot)$ and $\mathcal{N}^{O}_{Init}(\cdot)$. Both feature extractors do not share weights and are composed of three 3×3 convolutions with LeakyReLU, following the design used in most existing works (Liang et al. 2023, 2022; Wang et al. 2022c). The extracted initial features $\boldsymbol{F}^0$ and $\boldsymbol{F}^M$ are fed to the occlusion-embedded mix layer $\mathcal{N}_{mix}(\cdot)$, generating the occlusion-embedded feature representation $\boldsymbol{F}^1 \in \mathbb{R}^{H\times W\times U\times V\times C}$

$$\boldsymbol{F}^1 = \mathcal{N}_{mix}(\mathcal{N}^{LF}_{Init}(\boldsymbol{I}^{sr}), \mathcal{N}^{O}_{Init}(\boldsymbol{M})), \tag{2}$$

where C denotes the channel number. We set C to 48, 56, and 64, obtaining the tiny, small, and base versions of OHT, abbreviated as OHT-T, OHT-S, and OHT-B. Then, $\boldsymbol{F}^1$ is fed to N hybrid blocks, generating the HR feature representation $\boldsymbol{F}^N$. The hybrid blocks in our model are implemented following the design of EPIT (Liang et al. 2023), ensuring efficient spatial-angular feature learning. Finally, $\boldsymbol{F}^N$ is fed to the reconstructor $\mathcal{N}_{Rec}(\cdot)$. We obtain the super-resolved light field $\boldsymbol{I}^{sr}$ by adding the upsampled result of $\boldsymbol{I}^{lr}$ (*i.e.*, $\boldsymbol{I}^{bi}$) to the output of $\mathcal{N}_{Rec}$

$$\boldsymbol{I}^{sr} = \mathcal{N}_{Rec}(\boldsymbol{F}^N + \boldsymbol{F}^1) + \boldsymbol{I}^{bi}. \tag{3}$$

Occlusion Map Generator

Existing methods perform poorly in regions with occlusions, necessitating the calculation of an occlusion mask for each view to generate reasonable guidance for light field SR. However, accurate occlusion estimation is a non-trivial task. Inspired by unsupervised light field disparity estimation methods (Peng et al. 2020, 2018; Jin and Hou 2022; Wang et al. 2022b), we utilize a parameter-free approach to deduce the occlusion mask for each view.

Specifically, for regions with occlusions, a scene point visible in the center view may be unavailable in surrounding views, and the occluded pixels in these surrounding views cannot find their corresponding pixels in the center view. Consequently, the fine-grained occlusion mask can be calculated based on the photometric consistency prior. Denote the disparity map of the center view as $\mathcal{D}_c$. The surrounding views are first warped to the center view as follows

$$\mathcal{I}_{k\to c} = W^{\mathcal{D}_c}_{k\to c}(\mathcal{I}_k), \quad k = 1, 2, \cdots, UV, \tag{4}$$

where $W^{\mathcal{D}_c}_{k\to c}$ denotes the warping operation that projects the k-th view $\mathcal{I}_k$ to the center view $\mathcal{I}_c$. Assuming the disparity map $\mathcal{D}_c$ is accurate, the projected view $\mathcal{I}_{k\to c}$ should have identical values to the center view $\mathcal{I}_c$ at non-occluded regions. Therefore, we use the absolute residuals between $\mathcal{I}_{k\to c}$ and $\mathcal{I}_c$ to measure photometric consistency

$$\mathcal{I}^{res}_{k\to c} = |\mathcal{I}_{k\to c} - \mathcal{I}_c|. \tag{5}$$

Finally, the occlusion mask of the k-th view is obtained by re-mapping $\mathcal{I}^{res}_{k\to c}$ to $[0, 1]$

$$\mathcal{M}_k = |1 - \mathcal{I}^{res}_{k\to c}|^q, \tag{6}$$

where q is a scalar that controls the decaying rate. We empirically set $q = 2$ to achieve a good trade-off between occlusion awareness and noise robustness.

OACCNet constructs matching costs and handles occlusions by modulating input pixels. We adapt its structure for disparity estimation using a 5×5 angular resolution light field. With this modified OACCNet, we estimate disparity maps from the LR light field to generate occlusion maps. For more details, please refer to the supplementary document.

Occlusion-Embedded Mix Layer

Occlusions are prevalent in light fields and can deteriorate angular consistency. Proper utilization of occlusion information is essential. Existing methods often overlook the effect of occlusion during reconstruction, potentially leading to sub-optimal results. Therefore, in this paper, we propose to utilize occlusion maps to address these issues explicitly.

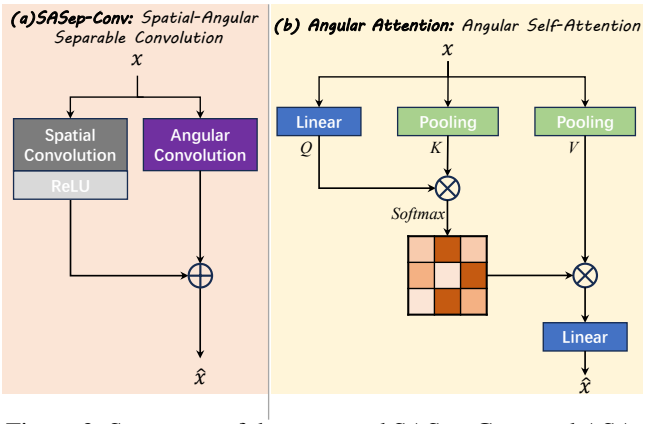

Figure 3: Structures of the proposed SASep-Conv and ASA.

The first step to fully leverage occlusion maps is to integrate occlusion information with light field data for subsequent reconstruction. We introduce a simple yet effective Occlusion-Embedded Mix Layer for this process. Specifically, we concatenate $\boldsymbol{F}^0$ and $\boldsymbol{F}^M$ along the feature dimension and pass them through the Occlusion-Embedded Mix Layer, which consists of four Conv3D layers, to obtain the fused feature $\boldsymbol{F}^0$.

Hybrid Transformer

In this section, we present a Hybrid Transformer, introducing several key designs, including (i) the SASep-Conv as an alternative to 3D convolution, (ii) the ASA, and (iii) the CFFL to enhance aggregated features.

SASep-Conv. Modern light field cameras capture high-dimensional images with numerous views. To process these complex data, 2D convolutions (Jin et al. 2020), 3D convolutions (Tran, Berberich, and Simon 2022) and 4D convolutions (Meng et al. 2019) have been employed. However, their performance and efficiency are constrained by the local receptive field and heavy computational demands. We introduce SASep-Conv, a lightweight and powerful variant of Conv3D that applies angular and spatial convolutions in parallel. This design preserves angular flexibility and augments a hybrid Transformer with inductive biases without splitting the light field into separate components. This enables more efficient processing and improved representation learning across both the angular and spatial domains.

Standard Conv3D effectively captures spectral-spatial correlations but introduces significant computational overhead due to its large parameter count. To address this, we propose SASep-Conv, which decouples Conv3D into two parallel branches that independently process spatial and angular dimensions. We first reshape the feature map from $H \times W \times U \times V \times C$ to $H \times W \times A \times C$, where $A = U \times V$. The spatial branch utilizes 2D filters per view, while the angular branch employs a 1×1 projection to capture angular correlations across all views. The final features are obtained by combining the outputs of both branches through element-wise addition. Unlike traditional 2D separable Conv for 3D data, SASep-Conv is spatial-angular separable rather than spatial-channel or angular-channel separable, making it effective for angular correlation extraction.

Angular self-attention. While spatial self-attention (Liang et al. 2021) enhances model performance by capturing spatial interactions and non-local similarities, it is computationally intensive and less suited for light field images. We introduce ASA, an efficient 3D self-attention mechanism that operates along the angular dimension rather than the spatial or channel dimensions. Leveraging the inherent angular correlations in light field data, the ASA mechanism achieves linear complexity and excels in long-range relation modeling. This design significantly enhances the model's ability to identify informative regions for SR, outperforming existing methods (Cong et al. 2023).

The proposed ASA mechanism operates on 3D feature maps generated by the preceding SASep-Conv, denoted as $\mathbf{X} \in \mathbb{R}^{H \times W \times D \times C}$. Unlike the traditional attention mechanism (Zamir et al. 2022), which applies attention along the channel dimension C, our ASA mechanism performs 3D attention along the angular dimension A, effectively capturing angular correlations within light field images.

To compute the attention, we first derive the query, key, and value from $\mathbf{X}$. However, unlike conventional attention mechanisms (Ashish 2017), where all three components undergo linear projection, in ASA, only the value $\mathbf{V} \in \mathbb{R}^{H \times W \times A \times C}$ is linearly projected from $\mathbf{X}$. To simplify the process, we omit intermediate operations and instead apply global average pooling to $\mathbf{X}$ along the spatial dimensions, directly obtaining the global features for each angular band, represented as $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{A \times C}$. This pooling strategy is both parameter-free and computationally efficient, avoiding the more considerable computational costs associated with the traditional reshape method used in prior works (Zamir et al. 2022).

The transposed attention map $\mathbf{A}$, with dimensions $\mathbb{R}^{D \times D}$, is then computed via the dot-product of key $\mathbf{K}$ and query $\mathbf{Q}$, followed by softmax normalization. The final attention output is obtained by multiplying this attention map with the value $\mathbf{V}$, allowing the model to dynamically focus on the most discriminative features across all views

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \text{Softmax}(\mathbf{K} \cdot \mathbf{Q}), \quad (7)$$

$$\hat{\mathbf{X}} = \mathbf{W} \cdot \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{X}. \quad (8)$$

To further refine the features, we apply an additional linear projection $\mathbf{W} \in \mathbb{R}^{C \times C}$ on the fused features, followed by the inclusion of residual connections to stabilize training. This design effectively leverages the most discriminative features for each view, enhancing the model's ability to process light field images.

Feed-forward layer. Feed-Forward Network (FFN) is one of the essential parts of Transformer architectures, and it has been reported that it might be the key to constructing the meta structure of transformer than SA (Yu et al. 2023). Traditional FFN (Ashish 2017) processes the output features from the SA layer with two linear projections and a nonlinear activation between them.

To effectively transform the features, we utilize CFFL, which consists of three convolutional layers and two LeakyReLU activation layers in between. This design enables the module to efficiently process and adaptively

Method	×2 SR					×4 SR				
	EPFL	HCInew	HCIold	INRIA	STFgantry	EPFL	HCInew	HCIold	INRIA	STFgantry
Bicubic	29.74/.9376	31.89/.9356	37.69/.9785	31.33/.9577	31.06/.9498	25.14/.8324	27.61/.8517	32.42/.9344	26.82/.8867	25.93/.8452
VDSR	32.50/.9598	34.37/.9561	40.61/.9867	34.43/.9741	35.54/.9789	27.25/.8777	29.31/.8823	34.81/.9515	29.19/.9204	28.51/.9009
EDSR	33.09/.9629	34.83/.9592	41.01/.9874	34.97/.9764	36.29/.9818	27.84/.8854	29.60/.8869	35.18/.9536	29.66/.9257	28.70/.9072
RCAN	33.16/.9634	34.98/.9603	41.05/.9875	35.01/.9769	36.33/.9831	27.88/.8863	29.63/.8886	35.20/.9548	29.76/.9276	28.90/.9131
ResLF	33.62/.9706	36.69/.9739	43.42/.9932	35.39/.9804	38.36/.9904	28.27/.9035	30.73/.9107	36.71/.9682	30.34/.9412	30.19/.9372
LFSSR	33.68/.9744	36.81/.9749	43.81/.9938	35.28/.9832	37.95/.9898	28.27/.9118	30.72/.9145	36.70/.9696	30.31/.9467	30.15/.9426
LF-ATO	34.27/.9757	37.24/.9767	44.20/.9942	36.15/.9842	39.64/.9929	28.52/.9115	30.88/.9135	37.00/.9699	30.71/.9484	30.61/.9430
LF-InterNet	34.14/.9760	37.28/.9763	44.45/.9946	35.80/.9843	38.72/.9909	28.67/.9162	30.98/.9161	37.11/.9716	30.64/.9491	30.53/.9409
DFNet	34.44/.9755	37.44/.9773	44.23/.9941	36.36/.9840	39.61/.9926	28.77/.9165	31.23/.9196	37.32/.9718	30.83/.9503	31.15/.9494
MEG-Net	34.30/.9773	37.42/.9777	44.08/.9942	36.09/.9849	38.77/.9915	28.74/.9160	31.10/.9177	37.28/.9716	30.66/.9490	30.77/.9453
LF-IINet	34.68/.9773	37.74/.9790	44.84/.9948	36.57/.9853	39.86/.9936	29.11/.9188	31.36/.9208	37.62/.9734	31.08/.9515	31.21/.9502
DPT	34.48/.9758	37.35/.9771	44.31/.9943	36.40/.9843	39.52/.9926	28.93/.9170	31.19/.9188	37.39/.9721	30.96/.9503	31.14/.9488
LFT	34.80/.9781	37.84/.9791	44.52/.9945	36.59/.9855	40.51/.9941	29.25/.9210	31.46/.9218	37.63/.9735	31.20/.9524	31.86/.9548
DistgSSR	34.81/.9787	37.96/.9796	44.94/.9949	36.59/.9859	40.40/.9942	28.99/.9195	31.38/.9217	37.56/.9732	30.99/.9519	31.65/.9535
LFSAV	34.62/.9772	37.43/.9776	44.22/.9942	36.36/.9849	38.69/.9914	*29.37*/*.9223*	31.45/.9217	37.50/.9721	31.27/*.9531*	31.36/.9505
EPIT	34.83/.9775	*38.23*/.9810	**45.08**/.9949	36.67/.9853	**42.17/.9957**	*29.34*/.9197	*31.51*/*.9231*	*37.68*/*.9737*	31.37/.9526	32.18/*.9571*
DET	35.26/.9797	38.31/*.9807*	*44.99*/**.9950**	36.95/.9864	41.76/.9955	**29.47**/.9230	31.56/.9235	37.84/.9744	**31.39**/.9534	*32.14*/.9573
OHT-T	34.89/*.9789*	38.03/.9800	44.86/*.9948*	36.48/.9857	40.28/.9940	29.17/.9181	31.30/.9202	37.35/.9720	31.15/.9504	31.33/.9511
OHT-S	*35.16*/.9797	38.21/.9805	44.98/.9949	*36.83*/*.9861*	40.99/.9949	29.25/.9197	31.42/.9216	37.48/.9727	31.16/.9514	31.51/.9525
OHT-B	**35.66/.9829**	**38.50/.9814**	45.04/**.9950**	**37.10/.9877**	*41.40*/*.9953*	29.45/**.9254**	**31.79/.9263**	**37.90/.9745**	*31.36*/**.9550**	**32.24/.9590**

Table 1: Quantitative comparison of different light field SR methods in terms of PSNR (dB) and SSIM on benchmark datasets. We mark the best, the second, and the third results in **bold**, underline, and *italic underlined*, respectively.

Method	Parameters		PSNR		SSIM	
	×2 SR	×4 SR	×2 SR	×4 SR	×2 SR	×4 SR
ResLF	6.35	6.79	37.49	31.24	.9817	.9321
LFSSR	0.81	1.61	37.50	31.52	.9832	.9370
LF-ATO	1.51	1.66	38.31	31.54	.9847	.9373
LF-InterNet	4.80	5.23	38.02	31.61	.9844	.9387
DFnet	3.94	3.99	38.53	31.92	.9854	.9422
MEG-Net	1.69	1.77	38.14	31.72	.9851	.9399
LF-IINet	4.84	4.89	38.76	32.06	.9860	.9429
DPT	3.73	3.78	38.40	31.93	.9848	.9414
LFT	1.11	1.16	38.83	32.27	.9861	.9445
DistgSSR	3.53	3.58	38.94	32.12	.9867	*.9439*
EPIT	1.42	1.47	*39.40*	*32.40*	.9869	.9452
DET	1.59	1.69	39.46	32.48	.9874	.9463
OHT-T	1.28	1.69	38.91	32.06	.9867	.9423
OHT-S	1.74	1.77	39.23	32.17	*.9872*	.9436
OHT-B	2.26	2.31	**39.54**	**32.55**	**.9885**	**.9480**

Table 2: Comparison in terms of parameters and average PSNR/SSIM values for ×2 and ×4 SR.

transform the input features, enhancing their discriminative power while maintaining computational efficiency.

Experiments

Experimental Settings

Datasets. In line with prior works, we utilize the BasicLFSR benchmark (Wang 2023) for training and evaluating all methods at ×2 and ×4 scales. This benchmark includes five light field datasets: HCI-new (Honauer et al. 2016), HCI-old (Wanner, Meister, and Goldluecke 2013), EPFL (Rerabek and Ebrahimi 2016), INRIA (Le Pendu, Jiang, and Guillemot 2018), and STFGantry, comprising 144 training scenes and 23 test scenes with diverse contents and disparities. We extract the central 5×5 views from each light field for both training and testing.

Training and inference settings. In the training stage, we crop each view into 32×32 or 64×64 patches and perform $\times 0.5$ or $\times 0.25$ bicubic downsampling to generate LR patches for ×2 and ×4 SR, respectively. We use PSNR and SSIM on the Y channel as quantitative metrics for performance evaluation. For a dataset with M scenes, we calculate metrics for each scene by averaging the scores of overall views and then obtain the dataset score by averaging the scores over M scenes.

Implementation details. We retrain the light field disparity estimation model at a 5×5 angular resolution following the settings of OACCNet. We then adopt the same training settings for all SR experiments, using the Xavier initialization algorithm and the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set $N = 6$. The initial learning rate is set to 2.5×10^{-4} and decreases by a factor of 0.8 every 15 epochs. The batch size is set to 128. During training, we perform random horizontal flipping, vertical flipping, and 90-degree rotation to augment the data. All models are implemented using the PyTorch framework. An NVIDIA A800 GPU is utilized for training. Specifically, the OHT-T and OHT-S models are trained from scratch for 75 epochs, while the OHT-B model is trained for 85 epochs.

Quantitative and Qualitative Comparisons

We compare our method with several baseline methods, including 3 single image SR methods (Kim, Lee, and Lee 2016; Lim et al. 2017; Zhang et al. 2018), and light field SR methods (Zhang, Lin, and Sheng 2019; Yeung et al. 2018; Wang et al. 2020, 2022c; Cheng, Liu, and Xiong 2022; Liang et al. 2022; Wang et al. 2022a; Cong et al. 2023; Liang et al. 2023) (see Table 1).

Quantitative results. The quantitative results are shown in Table 1 and Table 2. OHT-B achieves the highest PSNR and SSIM values in most cases, indicating superior image quality after SR. OHT-B uses 2.26M parameters for ×2 SR and 2.31M parameters for ×4 SR, which is more than OHT-T and OHT-S and baseline methods but still competitive. Compared to Transformer-based methods such as EPIT and DET, OHT models show competitive performance on most test sets. For instance, for the task of ×2 SR, OHT-B outper-

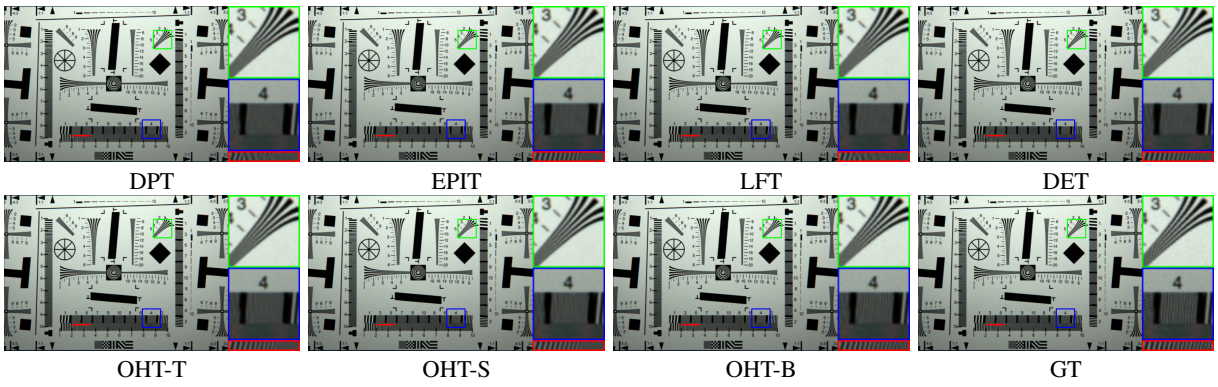

Figure 4: Qualitative comparison of different methods for $\times 2$ light field SR. Please zoom in for a better view.

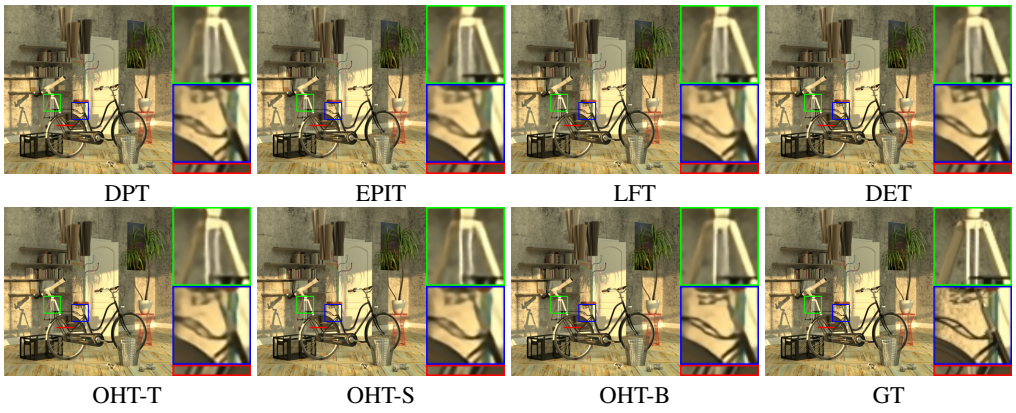

Figure 5: Qualitative comparison of different methods for $\times 4$ light field SR. Please zoom in for a better view.

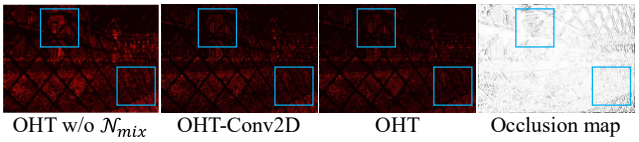

Figure 6: Ablations on the occlusion-embedded mix layer.

forms DET by 0.08 dB in PSNR.

Qualitative Results. We provide visual results on $\times 2$ SR and $\times 4$ SR in Figure 4 and Figure 5. For example, in the ISO scene in Fig. 4, OHT successfully generates precise results, while other baselines tend to generate blurry results and artifacts. More qualitative results are shown in the supplementary document.

Ablation Study

In this section, we perform ablation studies on $\times 4$ SR to validate the effectiveness of our designs based on OHT-B. We compare each component by replacing it with alternatives. When a component is removed, we maintain the parameter count by adding residual blocks. Due to space limitations, additional ablation study results and analyses are provided in the supplementary document.

Model	PSNR	SSIM
OHT w/o $\mathcal{N}_{mix}$	32.30	0.9451
$\mathcal{N}_{mix}$-Add	32.45	0.9459
$\mathcal{N}_{mix}$-Conv2D	32.48	0.9462
OHT	32.55	0.9480

Table 3: Comparison of different variants of the occlusion-embedded mix layer.

Model	PSNR	SSIM
Conv3D	32.41	0.9457
OHT w/o Spa	32.45	0.9458
OHT w/o Ang	32.47	0.9460
OHT	32.55	0.9480

Table 4: Comparison of different variants of SASep-Conv.

Effectiveness of the occlusion-embedded mix layer. $\mathcal{N}_{mix}$ integrates occlusion information with light field data for subsequent reconstruction. To validate this component, we design three variants: (1) OHT w/o $\mathcal{N}_{mix}$, where we remove $\mathcal{N}_{mix}$, meaning occlusions are not considered. (2) $\mathcal{N}_{mix}$-Add, where we add $\boldsymbol{F}^0$ and $\boldsymbol{F}^M$ before feeding them into $\mathcal{N}_{mix}$. (3) $\mathcal{N}_{mix}$-Conv2D, where we replace the 3D convolution with 2D convolution. The results are shown in Table 3.

Model	PSNR	SSIM
Channel Att.	32.39	0.9451
Restormer	32.42	0.9463
LFT	32.46	0.9468
EPIT	32.49	0.9470
OHT	32.55	0.9480

Table 5: Comparison of the angular self-attention.

Model	PSNR	SSIM
FFN	32.33	0.9455
OHT	32.55	0.9480

Table 6: Comparison of different variants of the convolutional feed-forward layer.

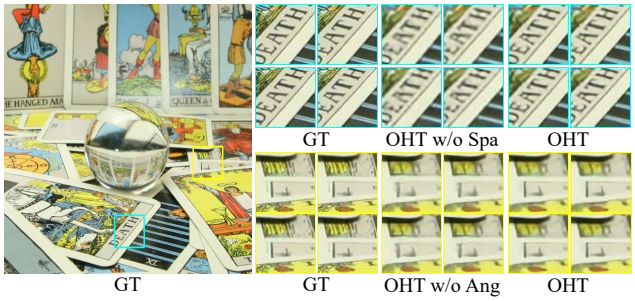

Figure 7: Ablations on the SASep-Conv.

The original OHT model outperforms the variant without the mixing network by 0.25 dB in PSNR and 0.0029 in SSIM, demonstrating the importance of considering occlusion information during SR. Our complete OHT, which incorporates the mixing network with 3D convolutions, yields the best performance, with a PSNR of 32.55 dB and an SSIM of 0.9480. Figure 6 showcases the visual comparison between the reconstructed images with and without explicit use of occlusion maps. It reveals that incorporating occlusion information in the reconstruction process leads to reduced errors, particularly in areas with heavy occlusions, emphasizing the importance of considering occlusion maps for enhanced light field SR.

Effectiveness of the SASep-Conv. To validate this component, we design three variants: (1) Conv3D: a baseline variant using only 3D convolutions to extract features. (2) OHT w/o Spa, OHT without the spatial branch. (3) OHT w/o Ang, OHT without the angular branch. The results are shown in Table 4. Combining both spatial and angular branches in the complete OHT results in the best performance, with a PSNR of 32.55 dB and an SSIM of 0.9480, outperforming Conv3D by 0.11 dB in PSNR and 0.0025 in SSIM. Figure 8 visually compares the reconstructed images obtained from OHT when removing either the spatial or angular branch. When the spatial branch is removed, noticeable artifacts appear, whereas eliminating the angular branch results in discontinuities and inconsistencies across different viewpoints.

Effectiveness of the ASA. To validate this component, we replace ASA with the attention modules listed in Table 4.

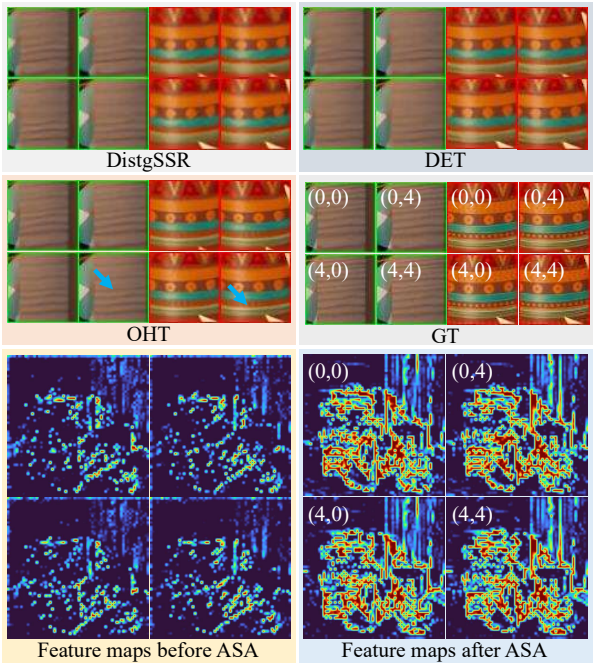

Figure 8: Ablations on the ASA mechanism.

Results in Table 4 suggest that ASA is more effective in capturing the unique characteristics of light fields compared to other techniques, leading to improved performance in light field SR. Figure 8 demonstrates the advantages of the ASA. The upper part of the figure shows that ASA produces angle-consistent results and effectively extracts useful texture information and repetitive patterns from different views for super-resolution. The lower part of the figure reveals that ASA activates more global features in the feature maps, highlighting its ability to capture long-range dependencies and enhance the representation power of the model.

Effectiveness of the CFFL. To evaluate the efficacy of CFFL, we substitute it with the FFN used in (Ashish 2017). The results presented in Table 6 indicate that CFFL outperforms FFN, demonstrating the superiority of CFFL.

Conclusion

In this paper, we propose a hybrid Transformer network for light field SR that efficiently captures both global and local spatial-angular correlations. By incorporating occlusion maps through an occlusion-embedded mix layer and leveraging SASep-Conv and ASA modules, our approach effectively preserves geometric details and enhances performance. Experimental results across multiple datasets demonstrate the superiority of our method over existing approaches, establishing a new benchmark for light field SR. Future work will focus on optimizing the model's complexity and exploring its applications in related tasks like light field image denoising and reflection removal.

Acknowledgments. We acknowledge funding from National Natural Science Foundation of China under Grants 62076086, 62476077 and 62272142.

References

Alain, M.; and Smolic, A. 2018. Light field super-resolution via LFBM5D sparse coding. In *IEEE Int. Conf. Image Process.*

Ashish, V. 2017. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30: I.

Chao, W.; Duan, F.; Wang, X.; Wang, Y.; and Wang, G. 2023. Lfsrdiff: Light field image super-resolution via diffusion models. *arXiv preprint arXiv:2311.16517*.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating more pixels in image super-resolution transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Cheng, Z.; Liu, Y.; and Xiong, Z. 2022. Spatial-angular versatile convolution for light field reconstruction. *IEEE Trans. Computational Imaging*, 8: 1131–1144.

Cheng, Z.; Xiong, Z.; and Liu, D. 2019. Light field super-resolution by jointly exploiting internal and external similarities. *IEEE Trans. Circ. Syst. Video Technol.*

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cong, R.; Sheng, H.; Yang, D.; Cui, Z.; and Chen, R. 2023. Exploiting Spatial and Angular Correlations With Deep Efficient Transformers for Light Field Image Super-Resolution. *IEEE Trans. Multimedia*.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Eur. Conf. Comput. Vis.*

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, R.; Liu, Y.; Xiao, Z.; and Xiong, Z. 2024. Diffusion-based Light Field Synthesis. *arXiv preprint arXiv:2402.00575*.

Gao, R.; Xiao, Z.; and Xiong, Z. 2024. Mamba-based light field super-resolution with efficient subspace scanning. In *Asian Conf. Comput. Vis.*, 531–547.

Honauer, K.; Johannsen, O.; Kondermann, D.; and Goldluecke, B. 2016. A dataset and evaluation methodology for depth estimation on 4D light fields. In *Asian Conf. Comput. Vis.*

Jin, J.; and Hou, J. 2022. Occlusion-aware unsupervised learning of depth from 4-d light fields. *IEEE Trans. Image Process.*, 31: 2216–2228.

Jin, J.; Hou, J.; Chen, J.; and Kwong, S. 2020. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Jin, K.; Yang, A.; Wei, Z.; Guo, S.; Gao, M.; and Zhou, X. 2023. Distgepit: Enhanced disparity learning for light field image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Le Pendu, M.; Jiang, X.; and Guillemot, C. 2018. Light field inpainting propagation via low rank matrix completion. *IEEE Trans. Image Process.*, 27(4): 1981–1993.

Levin, A.; Freeman, W. T.; and Durand, F. 2008. Understanding camera trade-offs through a Bayesian analysis of light field projections. In *Eur. Conf. Comput. Vis.*

Li, Z.; Li, J.; Li, Y.; Li, L.; Liu, D.; and Wu, F. 2024a. In-Loop Filtering via Trained Look-Up Tables. *arXiv preprint arXiv:2407.10926*.

Li, Z.; Li, Y.; Tang, C.; Li, L.; Liu, D.; and Wu, F. 2024b. Uniformly Accelerated Motion Model for Inter Prediction. *arXiv preprint arXiv:2407.11541*.

Li, Z.; Yuan, Z.; Li, L.; Liu, D.; Tang, X.; and Wu, F. 2024c. Object Segmentation-Assisted Inter Prediction for Versatile Video Coding. *arXiv preprint arXiv:2403.11694*.

Liang, C.-K.; and Ramamoorthi, R. 2015. A light transport framework for lenslet light field cameras. *ACM Trans. Graph.*, 34(2): 1–19.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Int. Conf. Comput. Vis.*

Liang, Z.; Wang, Y.; Wang, L.; Yang, J.; and Zhou, S. 2022. Light field image super-resolution with transformers. *IEEE Signal Process. Let.*, 29: 563–567.

Liang, Z.; Wang, Y.; Wang, L.; Yang, J.; Zhou, S.; and Guo, Y. 2023. Learning non-local spatial-angular correlation for light field image super-resolution. In *Int. Conf. Comput. Vis.*

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

Liu, G.; Yue, H.; and Yang, J. 2024. Efficient Light Field Image Super-Resolution via Progressive Disentangling. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Liu, Z.; Lin, Y.; Cao, Y.; and Hu, H. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*

Lu, Y.; Wang, S.; Wang, Z.; Xia, P.; Zhou, T.; et al. 2024. LFMamba: Light Field Image Super-Resolution with State Space Model. *arXiv preprint arXiv:2406.12463*.

Meng, N.; So, H. K.-H.; Sun, X.; and Lam, E. 2019. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*

Mi, Z.-Y.; and Yang, Y.-B. 2024. CUTDEM: Depth-Aware Enhanced Multi-View Image Mixing for Light Field Super-Resolution. In *IEEE Int. Conf. Acoust. Speech SP*. IEEE.

Ng, R.; Levoy, M.; Brédif, M.; Duval, G.; Horowitz, M.; and Hanrahan, P. 2005. *Light field photography with a hand-held plenoptic camera*. Ph.D. thesis, Stanford University.

Peng, J.; Xiong, Z.; Liu, D.; and Chen, X. 2018. Unsupervised depth estimation from light field using a convolutional neural network. In *Int. Conf. 3D Vis.*

Peng, J.; Xiong, Z.; Wang, Y.; Zhang, Y.; and Liu, D. 2020. Zero-shot depth estimation from light field using a convolutional neural network. *IEEE Trans. Computational Imaging*, 6: 682–696.

Raj, A. S.; Lowney, M.; Shah, R.; and Wetzstein, G. 2016. Stanford lytro light field archive.

Rerabek, M.; and Ebrahimi, T. 2016. New light field image dataset. In *QoMEX*.

Rossi, M.; and Frossard, P. 2017. Graph-based light field super-resolution. In *MMSP*.

Sheng, H.; Cong, R.; Yang, D.; Chen, R.; Wang, S.; and Cui, Z. 2022. UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Trans. Circ. Syst. Video Technol.*, 32(11): 7880–7893.

Tang, C.; Sheng, X.; Li, Z.; Zhang, H.; Li, L.; and Liu, D. 2024. Offline and Online Optical Flow Enhancement for Deep Video Compression. In *AAAI*, volume 38, 5118–5126.

Tian, C.; Zheng, M.; Zuo, W.; Zhang, S.; Zhang, Y.; and Lin, C.-W. 2024. A cross Transformer for image denoising. *Information Fusion*, 102: 102043.

Tran, T.-H.; Berberich, J.; and Simon, S. 2022. 3DVSR: 3D EPI volume-based approach for angular and spatial light field image super-resolution. *Signal Processing*, 192: 108373.

Van Duong, V.; Huu, T. N.; Yim, J.; and Jeon, B. 2023a. End-to-End Learned Light Field Image Rescaling Using Joint Spatial-Angular and Epipolar Information. In *IEEE Int. Conf. Image Process.*, 1935–1939.

Van Duong, V.; Huu, T. N.; Yim, J.; and Jeon, B. 2023b. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Trans. Computational Imaging*, 9: 350–366.

Wang, S.; Zhou, T.; Lu, Y.; and Di, H. 2022a. Detail preserving transformer for light field image super-resolution. In *AAAI*.

Wang, Y.; Liu, F.; Zhang, K.; Hou, G.; Sun, Z.; and Tan, T. 2018. LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Trans. Image Process.*, 27(9): 4274–4286.

Wang, Y.; Wang, L.; Liang, Z.; Yang, J.; An, W.; and Guo, Y. 2022b. Occlusion-aware cost constructor for light field depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Wang, Y.; Wang, L.; Wu, G.; Yang, J.; An, W.; Yu, J.; and Guo, Y. 2022c. Disentangling light fields for super-resolution and disparity estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*

Wang, Y.; Wang, L.; Yang, J.; An, W.; Yu, J.; and Guo, Y. 2020. Spatial-angular interaction for light field image super-resolution. In *Eur. Conf. Comput. Vis.*

Wang, Y. e. a. 2023. NTIRE 2023 Challenge on Light Field Image Super-Resolution: Dataset, Methods and Results. In *CVPRW*.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022d. Uformer: A general u-shaped transformer for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Wanner, S.; Meister, S.; and Goldluecke, B. 2013. Datasets and benchmarks for densely sampled 4D light fields. In *VMV*, volume 13, 225–226.

Wu, G.; Masia, B.; Jarabo, A.; Zhang, Y.; Wang, L.; Dai, Q.; Chai, T.; and Liu, Y. 2017. Light field image processing: An overview. *J. Sel. Top. Signal Process.*, 11(7): 926–954.

Xiao, Z.; Cheng, Z.; and Xiong, Z. 2023. Space-time super-resolution for light field videos. *IEEE Trans. Image Process.*

Xiao, Z.; Fu, X.; Huang, J.; Cheng, Z.; and Xiong, Z. 2021. Space-time distillation for video super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Xiao, Z.; Gao, R.; Liu, Y.; Zhang, Y.; and Xiong, Z. 2023a. Toward Real-World Light Field Super-Resolution. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

Xiao, Z.; Kai, D.; Zhang, Y.; Sun, X.; and Xiong, Z. 2024. Asymmetric Event-Guided Video Super-Resolution. In *ACMMM*.

Xiao, Z.; Liu, Y.; Gao, R.; and Xiong, Z. 2023b. CutMIB: Boosting Light Field Super-Resolution via Multi-View Image Blending. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Xiao, Z.; Shou, J.; and Xiong, Z. 2024. Learning Complementary Maps for Light Field Salient Object Detection. In *Asian Conf. Comput. Vis.*

Xiao, Z.; and Xiong, Z. 2024. Incorporating Degradation Estimation in Light Field Spatial Super-Resolution. *arXiv preprint arXiv:2405.07012*.

Yeung, H. W. F.; Hou, J.; Chen, X.; Chen, J.; Chen, Z.; and Chung, Y. Y. 2018. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Trans. Image Process.*, 28(5): 2319–2330.

Yoon, Y.; Jeon, H.-G.; Yoo, D.; Lee, J.-Y.; and Kweon, I. S. 2017. Light-field image super-resolution using convolutional neural network. *IEEE Signal Process. Let.*, 24(6): 848–852.

Yu, W.; Si, C.; Zhou, P.; Luo, M.; Zhou, Y.; Feng, J.; Yan, S.; and Wang, X. 2023. Metaformer baselines for vision. *IEEE Trans. Pattern Anal. Mach. Intell.*

Yuan, Y.; Cao, Z.; and Su, L. 2018. Light-field image super-resolution using a combined deep CNN based on EPI. *IEEE Signal Process. Let.*, 25(9): 1359–1363.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, S.; Chang, S.; and Lin, Y. 2021. End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Trans. Image Process.*, 30: 5956–5968.

Zhang, S.; Lin, Y.; and Sheng, H. 2019. Residual networks for light field image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*