Learning Non-Local Spatial-Angular Correlation for Light Field Image Super-Resolution

Zhengyu Liang[1], Yingqian Wang[1], Longguang Wang[2], Jungang Yang[1✉], Shilin Zhou[1], Yulan Guo[1]
[1]National University of Defense Technology, [2]Aviation University of Air Force
{zyliang, yangjungang}@nudt.edu.cn

Abstract

Exploiting spatial-angular correlation is crucial to light field (LF) image super-resolution (SR), but is highly challenging due to its non-local property caused by the disparities among LF images. Although many deep neural networks (DNNs) have been developed for LF image SR and achieved continuously improved performance, existing methods cannot well leverage the long-range spatial-angular correlation and thus suffer a significant performance drop when handling scenes with large disparity variations. In this paper, we propose a simple yet effective method to learn the non-local spatial-angular correlation for LF image SR. In our method, we adopt the epipolar plane image (EPI) representation to project the 4D spatial-angular correlation onto multiple 2D EPI planes, and then develop a Transformer network with repetitive self-attention operations to learn the spatial-angular correlation by modeling the dependencies between each pair of EPI pixels. Our method can fully incorporate the information from all angular views while achieving a global receptive field along the epipolar line. We conduct extensive experiments with insightful visualizations to validate the effectiveness of our method. Comparative results on five public datasets show that our method not only achieves state-of-the-art SR performance but also performs robust to disparity variations. Code is publicly available at https://github.com/ZhengyuLiang24/EPIT.

1. Introduction

Light field (LF) cameras record both intensity and directions of light rays, and enable various applications such as depth perception [25, 29, 32], view rendering [3, 52, 66], virtual reality [11, 74], and 3D reconstruction [6, 77]. However, due to the inherent spatial-angular trade-off [82], an LF camera can either provide dense angular samplings with low-resolution (LR) sub-aperture images (SAIs), or capture high-resolution (HR) SAIs with sparse angular sampling.

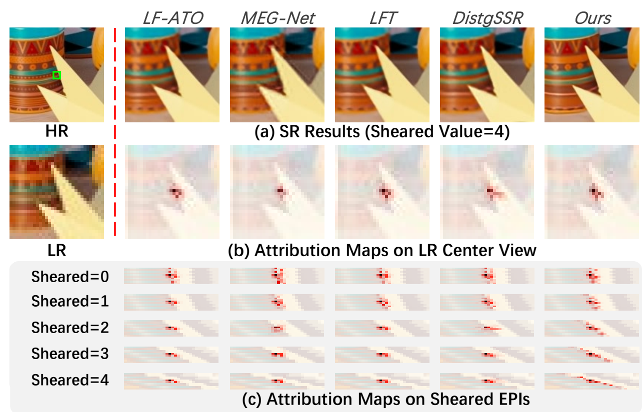

Figure 1. Visualization of $4\times$ SR results and the corresponding attribution maps [18] of our method and four state-of-the-art methods [26, 36, 60, 78] under different manually sheared disparity values. Here, the patch marked by the green box in the HR image is selected as the target region, and the regions that contribute to the final SR results are highlighted in red. Our method can well exploit the non-local spatial-angular correlation and achieve superior SR performance. More examples are provided in Fig. 8.

To handle this problem, many methods have been proposed to enhance the angular resolution via novel view synthesis [28, 43, 67], or enhance the spatial resolution by performing LF image super-resolution (SR) [10, 20]. In this paper, we focus on the latter task, i.e., generating HR LF images from their LR counterparts.

Recently, convolutional neural networks (CNNs) have been widely applied to LF image SR and demonstrated superior performance over traditional paradigms [1, 34, 44, 49, 64]. To incorporate the complementary information (i.e., angular information) from different views, existing CNNs adopted various mechanisms such as adjacent-view combination [73], view-stack integration [26, 78, 79], bidirectional recurrent fusion [59], spatial-angular disentanglement [36, 60, 61, 72, 9], and 4D convolutions [42, 43]. However, as illustrated in both Fig. 1 and Sec. 4.3, existing methods achieve promising results on LFs with small baselines, but suffer a notable performance drop when handling scenes with large disparity variations.

We attribute this performance drop to the contradictions between the local receptive field of CNNs and the requirement of incorporating *non-local spatial-angular correlation* in LF image SR. That is, LF images provide multiple observations of a scene from a number of regularly distributed viewpoints, and a scene point is projected onto different but correlated spatial locations on different angular views, which is termed as the *spatial-angular correlation*. Note that, the spatial-angular correlation has the non-local property since the difference between the spatial locations of two views (i.e., disparity value) depends on several factors (e.g., angular coordinates of the selected views, the depth value of the scene point, the baseline length of the LF camera, and the resolution of LF images), and can be very large in some situations. Consequently, it is appealing for LF image SR methods to incorporate complementary information from different views by exploiting the spatial-angular correlation under large disparity variations.

In this paper, we propose a simple yet effective method to learn the non-local spatial-angular correlation for LF image SR. In our method, we re-organize 4D LFs as multiple 2D epipolar plane images (EPIs) to manifest the spatial-angular correlation to the line patterns with different slopes. Then, we develop a Transformer-based network called EPIT to learn the spatial-angular correlation by modeling the dependencies between each pair of pixels on EPIs. Specifically, we design a basic Transformer block to alternately process horizontal and vertical EPIs, and thus progressively incorporate the complementary information from all angular views. Compared to existing LF image SR methods, our method can achieve a global receptive field along the epipolar line, and thus performs robust to disparity variations.

In summary, the contributions of this work are as follows: (1) We address the importance of exploiting non-local spatial-angular correlation in LF image SR, and propose a simple yet effective method to handle this problem. (2) We develop a Transformer-based network to learn the non-local spatial-angular correlation from horizontal and vertical EPIs, and validate the effectiveness of our method through extensive experiments and visualizations. (3) Compared to existing state-of-the-art LF image SR methods, our method achieves superior performance on public LF datasets, and is much more robust to disparity variations.

2. Related Work

2.1. LF Image SR

LFCNN [73] is the first method to adopt CNNs to learn the correspondence among stacked SAIs. Then, it is a common practice for LF image SR networks to aggregate the complementary information from adjacent views to model the correlation in LFs. Yeung et al. [72] designed a spatial-angular separable (SAS) convolution to approximate the 4D convolution to characterize the sub-pixel relationship of LF 4D structures. Wang et al. [59] proposed a bidirectional recurrent network to model the spatial correlation among views on horizontal and vertical baselines. Meng et al. [42] proposed a densely-connected network with 4D convolutions to explicitly learn the spatial-angular correlation encoded in 4D LF data. To further learn inherent corresponding relations in SAIs, Zhang et al. [78, 79] grouped LFs into four different branches according to the specific angular directions, and used four sub-networks to model the multi-directional spatial-angular correlation.

The aforementioned networks use part of input views to super-resolve each view, and the inherent spatial-angular correlation in LF images cannot be well incorporated. To address this issue, Jin et al. [26] proposed an All-to-One framework for LF image SR, and each SAI can be individually super-resolved by combining the information from all views. Wang et al. [61, 60] organized LF images into macro-pixels, and designed a disentangling mechanism to fully incorporate the angular information. Liu et al. [38] introduced 3D convolutions based multi-view context block to exploit the correlations among all views. In addition, Wang et al. [62] adopted deformable convolutions to achieve long-range information exploitation from all SAIs. Existing methods generally learn the local correspondence across SAIs, and ignore the importance of non-local spatial-angular correlation in LF images. However, due to the limited receptive field of CNNs, existing methods generally learn the local correspondence across SAIs, and ignore the importance of non-local spatial-angular correlation in LF images.

Recently, Liang et al. [36] applied Transformers to LF image SR and developed an angular Transformer and a spatial Transformer to incorporate angular information and model long-range spatial dependencies, respectively. However, since 4D LFs were organized into 2D angular patches to form the input of angular Transformers, the non-local property of spatial-angular correlations reduces the robustness of LFT to large disparity variations.

2.2. Non-Local Correlation Modeling

Non-local means [5] is a classical algorithm that computes the weighted mean of pixels in an image according to the self-similarity measure, and a number of studies on such non-local priors have been proposed for image restoration [12, 51, 19, 4], image and video SR [16, 76, 14, 71, 23]. Then, the attention mechanism is developed as a tool to bias the most informative components of an input signal, and achieves significant performance in various computer vision tasks [22, 8, 58, 15]. Huang et al. [24] proposed novel criss-cross attention to capture contextual information from full-image dependencies in an efficient way. Wang et al. [56, 55] proposed a parallax attention mechanism to handle the varying disparities problem of stereo images. Wu et al. [69] applied attention mechanisms to 3D LF reconstruction and developed a spatial-angular attention module to learn the first-order correlation on EPIs.

Recently, the attention mechanism is further generalized as Transformers [54] with multi-head structures and feed-forward networks. Transformers have inspired lots of works [39, 35, 7, 13] to further investigate the power of attention mechanisms in visions. Liu et al. [40] presented a pure-Transformer method to incorporate the inherent spatial-temporal locality of videos for action recognition. Naseer et al. [45] investigated the robustness and generalizability of Transformers, and demonstrated favorable merits of Transformers over CNNs for occlusion handling. Shi et al. [50] observed that Transformers can implicitly make accurate connections for misaligned pixels, and presented a new understanding of Transformers to process spatially unaligned images.

3. Method

3.1. Preliminary

Based on the two-plane LF parameterization model [33], an LF image is commonly formulated as a 4D function $\mathcal{L}(u,v,h,w) \in \mathbb{R}^{U\times V\times H\times W}$, where U and V represent angular dimensions, H and W represent spatial dimensions. The EPI sample of 4D LF is acquired with a fixed angular coordinate and a fixed spatial coordinate. Specifically, the horizontal EPI is obtained with constant u and h, and the vertical EPI is obtained with constant v and w.

As shown in Fig. 2, the EPIs not only record spatial structures at edges or textures, but also reflect the disparity information via line patterns of different slopes. Specifically, due to large disparities, the EPIs contain abundant spatial-angular correlation of LFs in a long-range way. Therefore, we propose to explore the non-local spatial-angular correlation from horizontal and vertical EPIs for LF image SR.

3.2. Network Design

As shown in Fig. 3(a), our network takes an LR LF $\mathcal{L}_{LR} \in \mathbb{R}^{U\times V\times H\times W}$ as its input, and produces an HR LF $\mathcal{L}_{SR} \in \mathbb{R}^{U\times V\times \alpha H\times \alpha W}$, where α presents the upscaling factor. Our network consists of three stages including initial feature extraction, deep spatial-angular correlation learning, and feature upsampling.

3.2.1 Initial Feature Extraction

As shown in Fig. 3(b), we follow most existing works [7, 35, 75] to use three 3×3 convolutions with LeakyReLU [41] as a *SpatialConv* layer to map each SAI to a high-dimensional feature. The initially extracted feature can be represented as $\boldsymbol{F} \in \mathbb{R}^{U\times V\times H\times W\times C}$, where C denotes the channel dimension.

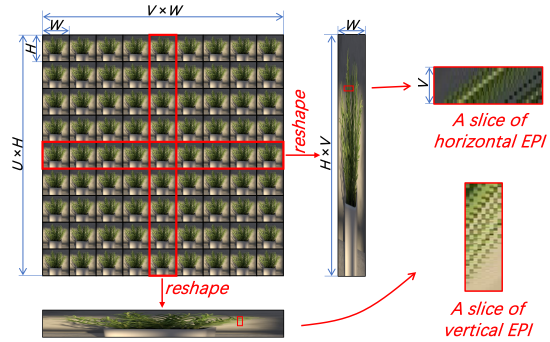

Figure 2. The SAI and EPI representations of LF images. The array of 9×9 views of a scene `rosemary` from HCInew [21] dataset is used as an example for illustration.

3.2.2 Deep Spatial-Angular Correlation Learning

Non-Local Cascading Block. The basic module for spatial-angular correlation learning is the *Non-Local Cascading* block. As shown in Fig. 3(a), each block consists of two cascaded *Basic-Transformer* units to separately incorporate the complementary information along the horizontal and vertical epipolar lines. In our method, we employed five *Non-Local Cascading* blocks to achieve a global perception of all angular views, and followed SwinIR [35] to adopt spatial convolutions to enhance the local feature representation. The effectiveness of this inter-block spatial convolution is validated in Sec. 4.4. Note that, the weights of the two *Basic-Transformer* units in each block are shared to jointly learn the intrinsic properties of LFs, which is demonstrated effective in Sec. 4.4.

As shown in Fig. 3(c), the initial features $\boldsymbol{F}$ can be first separately reshaped to the horizontal EPI pattern $\boldsymbol{F}_{hor} \in \mathbb{R}^{UH\times V\times W\times C}$ and the vertical EPI pattern $\boldsymbol{F}_{ver} \in \mathbb{R}^{VW\times U\times H\times C}$. Next, $\boldsymbol{F}_{hor}$ (or $\boldsymbol{F}_{ver}$) is fed to a *Basic-Transformer* unit to integrate the long-range information along the horizontal (or vertical) epipolar line. Then, the enhanced feature $\hat{\boldsymbol{F}}_{hor}$ (or $\hat{\boldsymbol{F}}_{ver}$) is reshaped into the size of $UV \times H \times W \times C$, and passes through a *SpatialConv* layer to incorporate the spatial context information within each SAI. Without loss of generality, we take the vertical *Basic-Transformer* as an example to introduce the detail of our *Basic-Transformer* unit in the following texts.

Basic-Transformer Unit. The objective of this unit is to capture long-range dependencies along the epipolar line via Transformers. To leverage the powerful sequence modeling capability of Transformers, we convert the vertical EPI features $\boldsymbol{F}_{ver}$ to the sequences of "tokens" for capturing spatial-angular correlation in U and H dimensions. As shown in Fig. 3(d), the vertical EPI features are passed through a linear projection matrix $W_{in} \in \mathbb{R}^{C\times D}$, where D

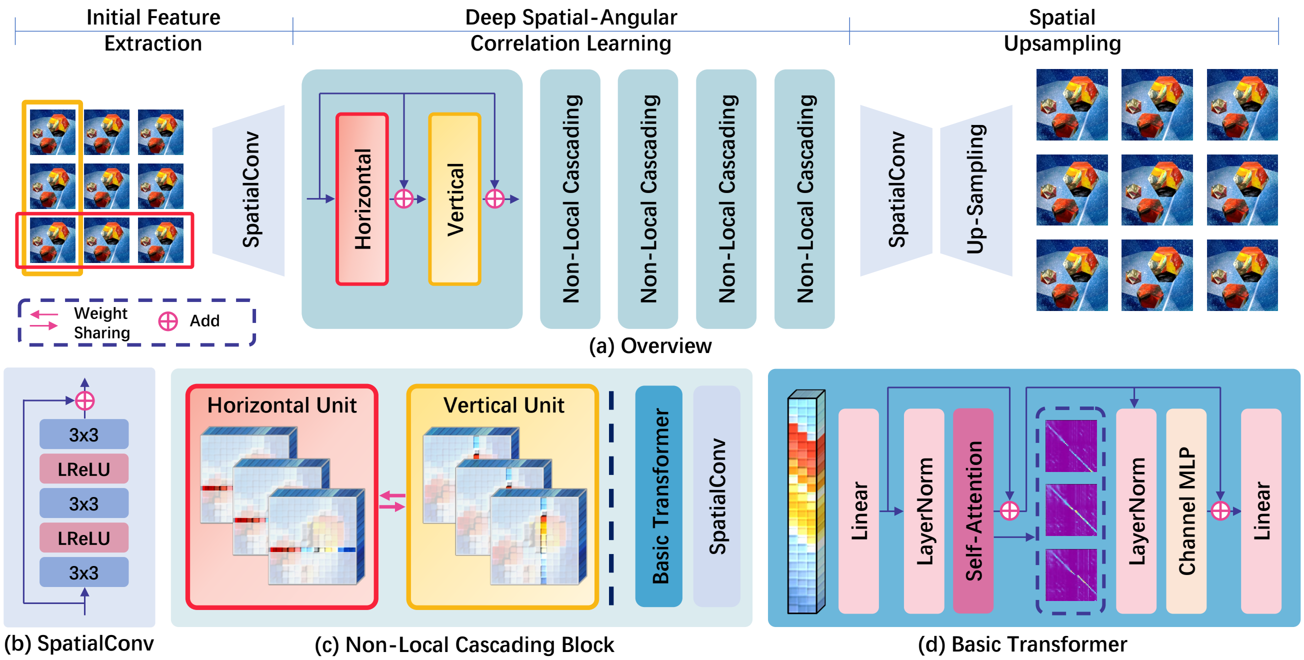

Figure 3. An overview of our proposed EPIT. Here, a 3×3 LF is used as an example for illustration.

denotes the embedding dimension of each token. The projected EPI features are a sequence of tokens with a length of UH, i.e., $\boldsymbol{T}_{ver} \in \mathbb{R}^{UH \times D}$. Following the PreNorm operation in [70], we also apply Layer Normalization (LN) before the attention calculation, and obtain the normalized tokens $\bar{\boldsymbol{T}}_{ver} = \mathrm{LN}(\boldsymbol{T}_{ver})$.

Afterwards, tokens $\bar{\boldsymbol{T}}_{ver}$ are passed through the *Self-Attention* layer and transformed into the deep tokens involving non-local spatial-angular information along the vertical epipolar line. Specifically, $\bar{\boldsymbol{T}}_{ver}$ need to be separately multiplied by $\boldsymbol{W}_Q \in \mathbb{R}^{D \times D}$, $\boldsymbol{W}_K \in \mathbb{R}^{D \times D}$ and $\boldsymbol{W}_V \in \mathbb{R}^{D \times D}$ to generate corresponding *query*, *key* and *value* components for self-attention calculation, i.e., $\boldsymbol{Q}_{ver} = \bar{\boldsymbol{T}}_{ver} W_Q$, $\boldsymbol{K}_{ver} = \bar{\boldsymbol{T}}_{ver} W_K$ and $\boldsymbol{V}_{ver} = \bar{\boldsymbol{T}}_{ver} W_V$.

Given a *query* position $q = \{1, 2, ..., UH\}$ in $\boldsymbol{Q}_{ver}$ and a *key* position $k = \{1, 2, ..., UH\}$ in $\boldsymbol{K}_{ver}$, the corresponding response $\boldsymbol{A}_{ver}(q, k) \in \mathbb{R}$ measures the mutual similarity of the pairs by the dot-product operation, followed by a Softmax function to obtain the attention scores on the vertical EPI tokens. That is,

$$\boldsymbol{A}_{ver}(q,k) = \mathrm{Softmax}(\frac{Q_{ver}(q) \cdot K_{ver}(k)^{\mathrm{T}}}{\sqrt{D}}). \tag{1}$$

Based on the attention scores, the output of self-attention $\boldsymbol{T}'_{ver}$ can be calculated as the weighted sum of *value*. In summary, the calculation process of *Self-Attention* layer can be formulated as:

$$\boldsymbol{T}'_{ver} = \boldsymbol{A}_{ver} V_{ver} + T_{ver}. \tag{2}$$

To further incorporate the spatial-angular correlation,

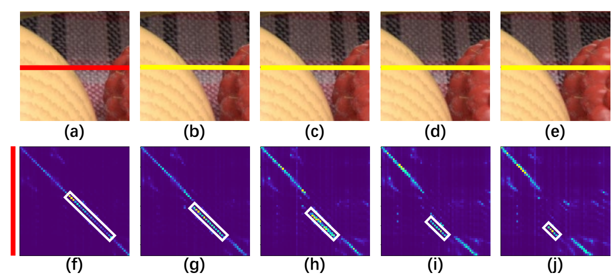

Figure 4. An example of the attention maps of a *Basic-Transformer* unit for the spatial-angular correlation. Note that, the attention maps correspond to the correlation between the regions marked by the red and yellow strokes.

following [54], our *Basic-Transformer* unit also contains the multi-layer perception (MLP) and LN, and generates the enhanced $\hat{\boldsymbol{T}}_{ver}$ as:

$$\hat{\boldsymbol{T}}_{ver} = \mathrm{MLP}(\mathrm{LN}(\boldsymbol{T}'_{ver})) + \boldsymbol{T}'_{ver}. \tag{3}$$

At the end of the *Basic-Transformer* unit, the enhanced $\hat{\boldsymbol{T}}_{ver}$ is fed to another linear projection $\boldsymbol{W}_{out} \in \mathbb{R}^{D \times C}$, and reshaped into the size of $UV \times H \times W \times C$ for the subsequent *SpatialConv* layer.

Cross-View Similarity Analysis. Note that, the setting $[\boldsymbol{A}_{ver}(q, 1), ..., \boldsymbol{A}_{ver}(q, UH)] \in \mathbb{R}^{1 \times UH}$ represents the similarity scores of q with all k in K_{ver}, and thus can be re-organized as a slice of cross-view attention map according to the angular coordinate. Inspired by this, we visualized the cross-view attention maps on an example scene in Fig. 4. The regions marked by the red stripe in Fig. 4(a) are set as the *query* tokens, and the self-similarity (i.e.,

Table 1. Quantitative comparison of different SR methods in terms of the number of parameters (#Prm.) and PSNR/SSIM. Larger PSNR and SSIM values indicate higher SR quality. We mark the best results in red and the second best results in blue.

Methods	#Prm.(M)	2×					4×				
	2×/4×	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 [30]	0.66 / 0.66	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 [37]	38.6 / 38.9	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 [81]	15.3 / 15.4	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[79]	7.98 / 8.64	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 [72]	0.88 / 1.77	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 [26]	1.22 / 1.36	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 [61]	5.04 / 5.48	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
LF-DFnet [62]	3.94 / 3.99	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 [78]	1.69 / 1.77	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 [38]	4.84 / 4.88	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 [57]	3.73 / 3.78	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 [36]	1.11 / 1.16	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 [60]	3.53 / 3.58	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 [9]	1.22 / 1.54	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 (ours)	1.42 / 1.47	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

key are same as *query*) is ideally located at the diagonal, as shown in Fig. 4(f). In contrast, the yellow stripes in Figs. 4(b)-4(e) are set as the *key* tokens, the corresponding cross-view similarities are shown in Figs. 4(g)-4(j). It can be observed that due to the foreground occlusions, the responses of the background appear as short lines (marked by the white boxes) parallel to the diagonal in each cross-view attention map, and both of the distance to the diagonal and the length of response regions adaptively change as the *key* view moves along the baseline, which demonstrates the disparity-awareness of our *Basic-Transformer* unit.

3.2.3 Feature Upsampling

Finally, we apply the pixel shuffling operation to increase the spatial resolution of LF features, and further employ a 3×3 convolution to obtain the super-resolved LF image $\mathcal{L}_{SR}$. Following most existing works [61, 60, 36, 62, 57, 38, 78, 79, 72], we use the L_1 loss function to train our network due to its robustness to outliers [2].

4. Experiments

In this section, we first introduce the datasets and our implementation details, and then compare our method with state-of-the-art methods. Next, we investigate the performance of different SR methods with respect to disparity variations. Finally, we validate the effectiveness of our method through ablation studies.

4.1. Datasets and Implementation Details

We followed [62, 60, 38, 57, 36] to use five public LF datasets (EPFL [48], HCInew [21], HCIold [65], INRIA [46], STFgantry [53]) in the experiments. All LFs in these datasets have an angular resolution of 9×9. Unless specifically mentioned, we extracted the central 5×5 SAIs for training and test. In the training stage, we cropped each SAI into patches of size 64×64/128×128, and performed $0.5\times/0.25\times$ bicubic downsampling to generate the LR patches for $2\times/4\times$ SR, respectively. We used peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [63] as quantitative metrics for performance evaluation. To obtain the metric score for a dataset with M scenes, we first calculated the metric of each scene by averaging the scores over all the SAIs separately, and then obtained the score for this dataset by averaging the scores over the M scenes.

We adopted the same training settings for all experiments, i.e., Xavier initialization algorithm [17] and Adam optimizer [31] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate was set to 2×10^{-4} and decreased by a factor of 0.5 every 15 epochs. During the training phase, we performed random horizontal flipping, vertical flipping, and 90-degree rotation to augment the training data. All models were implemented in the PyTorch framework and trained from scratch for 80 epochs with 2 Nvidia RTX 2080Ti GPUs.

4.2. Comparisons on Benchmark Datasets

We compare our method to 14 state-of-the-art methods, including 3 single image SR methods [30, 37, 81] and 11 LF image SR methods [79, 72, 26, 61, 62, 78, 38, 57, 36, 60, 9].

Quantitative Results. A quantitative comparison among different methods is shown in Tabel 1. Our EPIT with a small model size (i.e., 1.42M/1.47M for $2\times/4\times$ SR) achieves state-of-the-art PSNR and SSIM scores on almost all the datasets for both 2× and 4× SR. It is worth noting that LFs in the STFgantry dataset [53] have larger disparity variations, and are thus more challenging. Our EPIT significantly outperforms all the compared methods and achieves 1.66dB/0.32dB PSNR improvements over the second top-performing method LFT for $2\times/4\times$ SR, respectively, which demonstrates the powerful capacity of our EPIT in non-local correlation modeling.

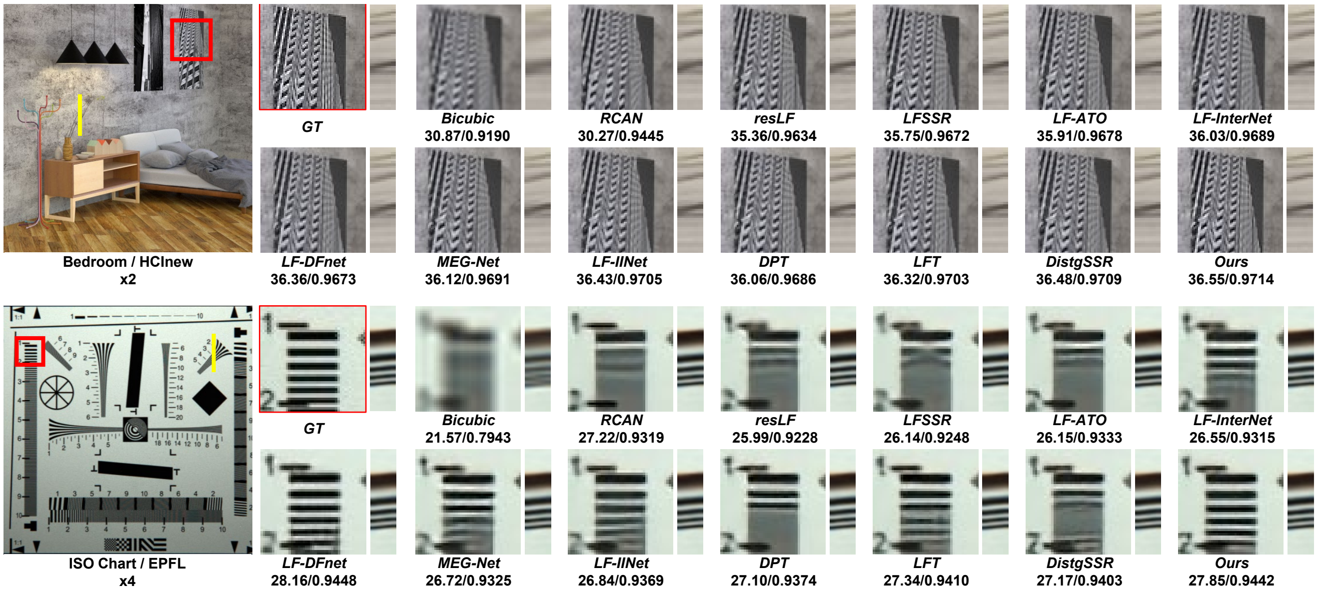

Figure 5. Qualitative comparison of different SR methods for 2×/4× SR.

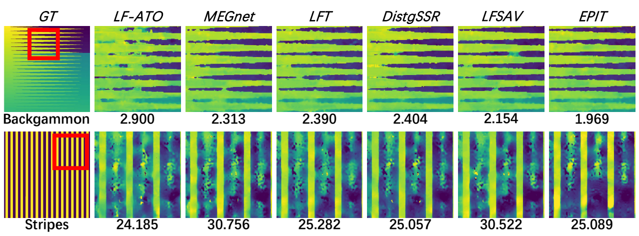

Figure 6. Quantitative and qualitative (MSE) comparisons of disparity estimation results achieved by SPO [80] using different SR results. The MSE (↓) is the mean square error multiplied by 100.

Table 2. PSNR values achieved by DistgSSR [60] and our EPIT with different angular resolution for 4× SR.

Input	EPFL		HCInew		HCIold		INRIA		STFgantry	
	[60]	Ours	[60]	Ours	[60]	Ours	[60]	Ours	[60]	Ours
2×2	28.27	-0.05	30.80	+0.04	36.77	+0.17	30.55	-0.03	30.74	+0.56
3×3	28.67	+0.03	31.07	+0.19	37.18	+0.19	30.83	+0.11	31.12	+0.74
4×4	28.81	+0.23	31.25	+0.15	37.32	+0.20	30.93	+0.26	31.23	+0.88
5×5	28.99	+0.35	31.38	+0.13	37.56	+0.12	30.99	+0.38	31.65	+0.56
6×6	29.10	+0.33	31.39	+0.18	37.52	+0.26	30.98	+0.47	31.57	+0.74
7×7	29.38	+0.22	31.43	+0.20	37.65	+0.27	31.18	+0.33	31.63	+0.77
8×8	29.32	+0.28	31.52	+0.14	37.76	+0.24	31.23	+0.31	31.58	+0.90
9×9	29.41	+0.30	31.48	+0.21	37.80	+0.26	31.22	+0.34	31.66	+0.84

Qualitative Results. Figure 5 shows the qualitative results achieved by different methods for 2×/4× SR. It can be observed from the zoom-in regions that single image SR method RCAN [81] cannot recover the textures and details in the SR images. In contrast, our EPIT can incorporate sub-pixel correspondence among SAIs and generate more faithful details with fewer artifacts. Compared to most LF image SR methods, our EPIT can generate superior visual results with high angular consistency. Please refer to the supplemental material for additional visual comparisons.

Angular Consistency. We evaluate the angular consistency by using the 4× SR results on several challenging scenes

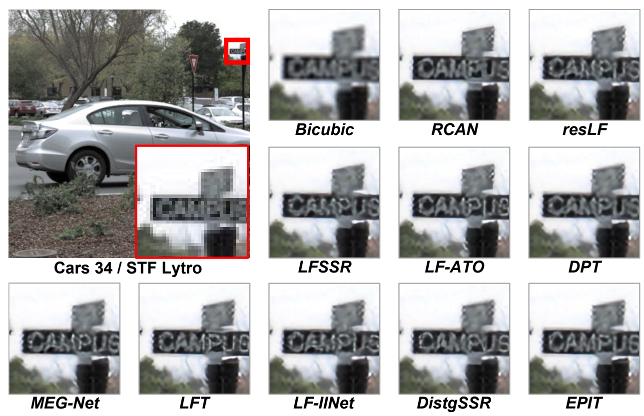

Figure 7. Visual comparison of different SR methods on real-world LF scenes for 4× SR.

(e.g., `Backgammon` and `Stripes`) in 4D LF benchmark [21] for disparity estimation. As shown in Fig. 6, our EPIT achieves competitive MSE scores on these challenging scenes, which demonstrates the superiority of our EPIT on angular consistency.

Performance with Different Angular Resolution. Since the angular resolution of LR images can vary significantly with different LF devices, we compare our method to DistgSSR [60] on LFs with different angular resolutions. It can be observed from Table 2 that our method achieves higher PSNR values than DistgSSR on almost all the datasets with each angular resolution (except on the EPFL and INRIA datasets with 2×2 input LFs). The consistent performance improvements demonstrate that our EPIT can well model the spatial-angular correlation with various angular resolutions. More comparisons and discussions are provided in the supplemental material.

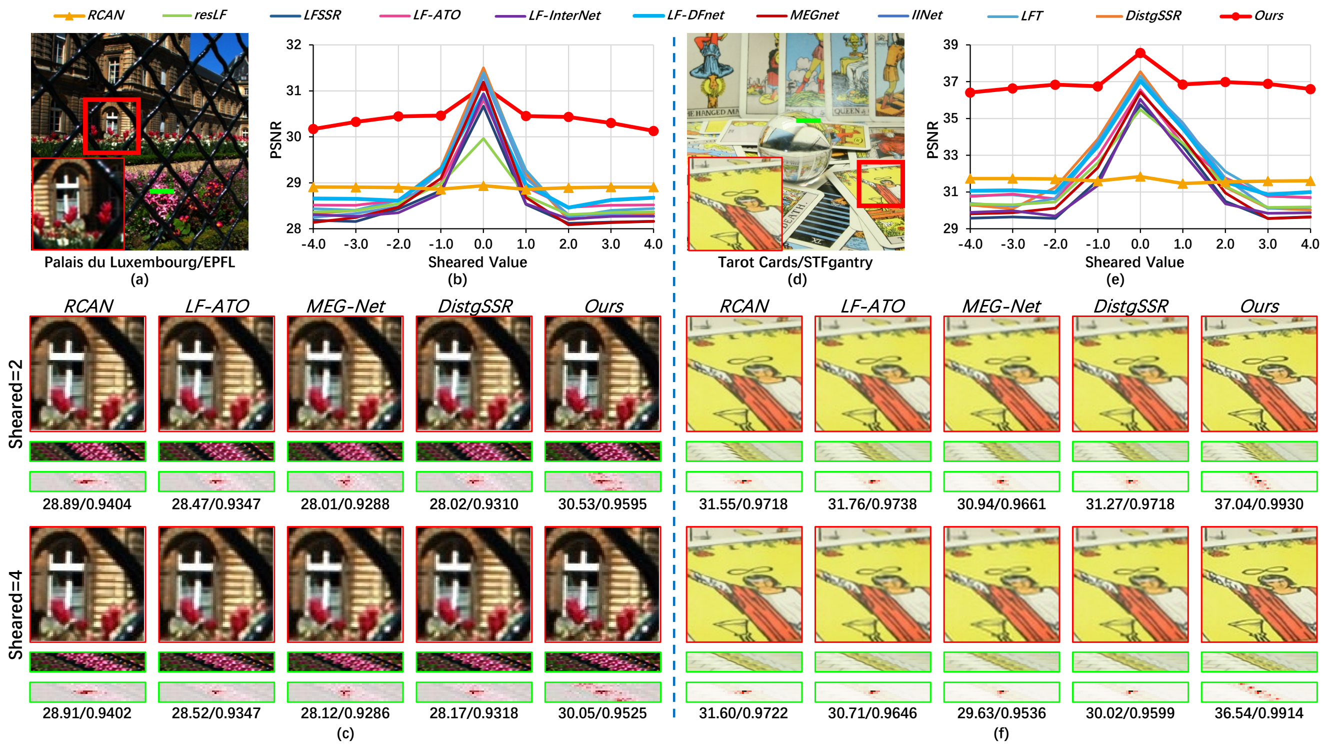

Figure 8. Performance comparison and local attribution maps of different SR methods on two representative scenes with different shearing values for $2\times$ SR. Here, we plot the performance curve to quantitatively measure the effect of disparity variations on LFs, and present the visual results and corresponding attribution maps under sheared value=2, 4.

Performance on Real-World LF Scenes. We compare our method to state-of-the-art methods under real-world degradation by directly applying them to LFs in the STFlytro dataset [47]. Since no groundtruth HR images are available in this dataset, we present the LR input and their super-resolved results in Fig. 7. It can be observed that our method can recover more faithful details and generate more clear letters than other methods. Since the LF structure keeps unchanged under both bicubic and real-world degradation, our method can learn the spatial-angular correlation from bicubicly downsampled training data, and well generalize to LF images under real degradation.

4.3. Robustness to Large Disparity Variations

Considering the parallax structure of LF images, we followed the shearing operation in existing works [67, 68] to linearly change the overall disparity range of LF datasets. Note that, the content of SAIs maintain unchanged after the shearing operation, and thus we can quantitatively investigate the performance of different SR methods with respect to the disparity variations.

Quantitative & Qualitative Comparison. Figure 8 shows the quantitative and qualitative results of different SR methods with respect to sheared values, from which we can observe that: 1) Except for the single image SR method RCAN, all LF image SR methods suffer a performance drop when the absolute sheared value of LF images increases. That is because, large sheared values can result in more significant misalignment among LF images, and introduce difficulties in complementary information incorporation; 2) As the absolute sheared value increases, the performance of existing LF image SR methods is even inferior to RCAN. The possible reason is that, these methods do not make full use of local spatial information, but rather rely on local angular information from adjacent views. When the sheared value exceeds their receptive fields, the large disparities can make the spatial-angular correlation non-local and thus introduce challenges in complementary information incorporation; 3) Our EPIT performs much more robust to disparity variations and achieves the highest PSNR scores under all sheared values. More quantitative comparisons on the whole datasets can be referred to the supplemental material.

LAM Visualization. We used Local Attribution Map (LAM) [18] to visualize the input regions that contribute to the SR results of different methods. As shown in Fig. 8, we first specify the center of green stripes in HR images as the target regions, and then re-organize the corresponding attribution maps on LR images into the EPI patterns. It can be observed that RCAN achieves a larger receptive field along the spatial dimension than other compared methods, which supports the results in Figs. 8(b) and 8(e) that RCAN achieves a relatively stable SR performance with different

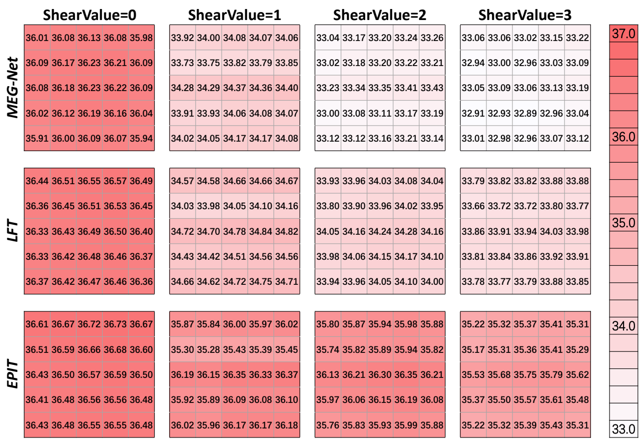

Figure 9. PSNR distribution among different SAIs achieved by MEG-Net [78], LFT [36], and our EPIT on the INRIA dataset [46] for 2× SR.

sheared values. It is worth noting that our EPIT can automatically incorporate the most relevant information from different views, and can learn the non-local spatial-angular correlation regardless of disparity variations.

Perspective Comparison. We compare the performance of MEG-Net, DistgSSR and our method with respect to different perspectives and sheared values (0 to 3). It can be observed in Fig. 9 that, both MEG-Net and DistgSSR suffer significant performance drops on all perspectives as the sheared value increases. In contrast, our EPIT can well handle the disparity variation problem, and achieve much higher PSNR values with a balanced distribution among different views regardless of the sheared values.

4.4. Ablation Study

In this subsection, we compare the performance of our EPIT with different variants to verify the effectiveness of our design choices, and additionally, investigate their robustness to large disparity variations.

Horizontal/Vertical Basic-Transformer Units. We demonstrated the effectiveness of the horizontal and vertical *Basic-Transformer* units in our EPIT by separately removing them from our network. Note that, without using horizontal or vertical *Basic-Transformer* unit, these variants cannot incorporate any information from the corresponding angular directions. As shown in Table 3, both variants *w/o-Horizontal* and *w/o-Vertical* suffer a decrease of 0.72dB in the INRIA dataset as compared to EPIT, which demonstrates the importance of exploiting spatial-angular correlations from all angular views.

Weight Sharing in Non-Local Cascading Blocks. We introduced the variant *w/o-Share* by removing the weight sharing between horizontal and vertical *Basic-Transformer* units. As shown in Table 3, the additional parameters in variant *w/o-Share* do not introduce further performance improvement. It demonstrates that the weight sharing strategy between two directional *Basic-Transformer* units is beneficial and efficient to regularize the network.

Table 3. The PSNR scores achieved by different variants of our EPIT on the LFs with different shearing values for 2× SR. We adjusted the channel number of each variant to make its model size (i.e., #Prm.) not smaller than EPIT for better validation.

Variants	#Prm.	FLOPs	EPFL (Sheared)			INRIA (Sheared)		
			0	2	4	0	2	4
w/o-Horiz	1.42M	80.20G	33.96	33.98	34.02	35.95	36.08	36.11
w/o-Verti	1.42M	80.20G	34.01	33.94	33.87	35.95	35.97	36.02
w/o-Share	2.71M	80.20G	34.80	34.63	34.51	36.66	36.72	36.45
w/o-Local	1.64M	96.39G	34.42	34.36	34.27	36.36	36.40	36.25
w/o-Trans	1.60M	78.82G	33.90	31.32	31.74	35.95	33.28	33.55
w-1-Block	1.54M	68.23G	33.97	34.24	34.08	35.84	36.19	35.93
w-2-Block	1.45M	73.37G	34.19	34.36	34.29	35.98	36.27	35.99
w-3-Block	1.71M	85.78G	34.64	34.51	34.45	36.53	36.47	36.22
EPIT	1.42M	74.96G	34.83	34.69	34.59	36.67	36.75	36.59

SpatialConv in Non-Local Cascading Blocks. We introduced the variant *w/o-Local* by removing the *SpatialConv* layers from our EPIT, and we adjusted the channel number to make the model size of this variant not smaller than the main model. As shown in Table 3, the *SpatialConv* has a significant influence on the SR performance, e.g., the variant *w/o-Local* suffers a 0.41dB PSNR drop on the EPFL dataset. It demonstrates that local context information is crucial to the SR performance, and the simple convolutions can fully incorporate the spatial information from each SAI.

Basic-Transformer in Non-Local Cascading Blocks. We introduced the variant *w/o-Trans* by replacing Basic-Transformer in Non-Local Blocks with cascaded convolutions. As shown in Table 3, *w/o-Trans* suffers a most significant performance drop as the sheared value increases, which demonstrates the effectiveness of the Basic-Transformer in incorporating global information on the EPIs.

Basic-Transformer Number. We introduced the variants *with-n-Block* (n=1,2,3) by retaining n Non-Local Blocks. Results in Table 3 show the effectiveness of our EPIT (having 5 Non-Local Blocks) with higher-order spatial-angular correlation modeling capability.

5. Conclusion

In this paper, we propose a Transformer-based network for LF image SR. By modeling the dependencies between each pair of pixels on EPIs, our method can learn the spatial-angular correlation while achieving a global receptive field along the epipolar line. Extensive experimental results demonstrated that our method can not only achieve state-of-the-art SR performance on benchmark datasets, but also perform robust to large disparity variations.

Acknowledgment: This work was supported in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61921001.

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, 2018.

[2] Yildiray Anagun, Sahin Isik, and Erol Seke. Srlibrary: comparing different loss functions for super-resolution over various convolutional architectures. *Journal of Visual Communication and Image Representation*, 61:178–187, 2019.

[3] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19819–19829, 2022.

[4] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1674–1682, 2016.

[5] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 60–65, 2005.

[6] Zewei Cai, Xiaoli Liu, Xiang Peng, and Bruce Z Gao. Ray calibration and phase mapping for structured-light-field 3d reconstruction. *Optics Express*, 26(6):7598–7613, 2018.

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12310, 2021.

[8] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3640–3649, 2016.

[9] Zhen Cheng, Yutong Liu, and Zhiwei Xiong. Spatial-angular versatile convolution for light field reconstruction. *IEEE Transactions on Computational Imaging*, 8:1131–1144, 2022.

[10] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10010–10019, 2021.

[11] Suyeon Choi, Manu Gopakumar, Yifan Peng, Jonghyun Kim, and Gordon Wetzstein. Neural 3d holography: Learning accurate wave propagation models for 3d holographic virtual and augmented reality displays. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021.

[12] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning and Representation (ICLR)*, 2015.

[14] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011.

[15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[16] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 349–356, 2009.

[17] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[18] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9199–9208, 2021.

[19] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2862–2869, 2014.

[20] Mantang Guo, Junhui Hou, Jing Jin, Jie Chen, and Lap-Pui Chau. Deep spatial-angular regularization for light field imaging, denoising, and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6094–6110, 2021.

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

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[23] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015.

[24] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 603–612, 2019.

[25] Jing Jin and Junhui Hou. Occlusion-aware unsupervised learning of depth from 4-d light fields. *IEEE Transactions on Image Processing*, 31:2216–2228, 2022.

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

[27] Jing Jin, Junhui Hou, Jie Chen, Huanqiang Zeng, Sam Kwong, and Jingyi Yu. Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[28] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–10, 2016.

[29] Numair Khan, Min H Kim, and James Tompkin. Differentiable diffusion for dense depth estimation from multi-view images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8912–8921, 2021.

[30] Jiwon Kim, JungKwon Lee, and KyoungMu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[31] DiederikP Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning and Representation (ICLR)*, 2015.

[32] Titus Leistner, Radek Mackowiak, Lynton Ardizzone, Ullrich Köthe, and Carsten Rother. Towards multimodal depth estimation from light fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12953–12961, 2022.

[33] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.

[34] Chia-Kai Liang and Ravi Ramamoorthi. A light transport framework for lenslet light field cameras. *ACM Transactions on Graphics (TOG)*, 34(2):1–19, 2015.

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021.

[36] 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.

[37] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and KyoungMu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144, 2017.

[38] Gaosheng Liu, Huanjing Yue, Jiamin Wu, and Jingyu Yang. Intra-inter view interaction network for light field image super-resolution. *IEEE Transactions on Multimedia*, pages 1–1, 2021.

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[40] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022.

[41] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[42] Nan Meng, HaydenKwokHay So, Xing Sun, and Edmund Lam. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[43] Nan Meng, Xiaofei Wu, Jianzhuang Liu, and Edmund Lam. High-order residual network for light field super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11757–11764, 2020.

[44] Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 22–28, 2012.

[45] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[46] MikaelLe 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.

[47] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. Stanford lytro light field archive, 2016.

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

[49] 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.

[50] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *Advances in Neural Information Processing Systems*, 2022.

[51] Abhishek Singh, Fatih Porikli, and Narendra Ahuja. Super-resolving noisy images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2853, 2014.

[52] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8269–8279, 2022.

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

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[55] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[56] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[57] 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,*, 2022.

[58] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

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

[60] 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*, 2022.

[61] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 290–308, 2020.

[62] 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.

[63] Zhou Wang, AlanC Bovik, HamidR Sheikh, and EeroP Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[64] 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.

[65] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization (VMV)*, volume 13, pages 225–226, 2013.

[66] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8534–8543, 2021.

[67] Gaochang Wu, Yebin Liu, Qionghai Dai, and Tianyou Chai. Learning sheared epi structure for light field reconstruction. *IEEE Transactions on Image Processing*, 28(7):3261–3273, 2019.

[68] Gaochang Wu, Yebin Liu, Lu Fang, and Tianyou Chai. Revisiting light field rendering with deep anti-aliasing neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[69] Gaochang Wu, Yingqian Wang, Yebin Liu, Lu Fang, and Tianyou Chai. Spatial-angular attention network for light field reconstruction. *IEEE Transactions on Image Processing*, 30:8999–9013, 2021.

[70] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533, 2020.

[71] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1059–1066, 2013.

[72] HenryWingFung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and YukYing Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018.

[73] Youngjin Yoon, HaeGon Jeon, Donggeun Yoo, JoonYoung Lee, and InSo Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24(6):848–852, 2017.

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

[75] Jiyang Yu, Jingen Liu, Liefeng Bo, and Tao Mei. Memory-augmented non-local attention for video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17834–17843, 2022.

[76] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730, 2010.

[77] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6525–6534, 2021.

[78] 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.

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

[80] 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.

[81] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[82] Hao Zhu, Mantang Guo, Hongdong Li, Qing Wang, and Antonio Robles-Kelly. Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution. *IEEE Transactions on Visualization and Computer Graphics*, 27(6):3019–3033, 2019.