DistgEPIT: Enhanced Disparity Learning for Light Field Image Super-Resolution

Kai Jin[1*] Angulia Yang[1*] Zeqiang Wei[3*] Sha Guo[2] Mingzhi Gao[1] Xiuzhuang Zhou[3†]

Bigo Technology Pte. Ltd.[1] Institute of Digital Media, Peking University[2]
School of Artificial Intelligence, Beijing University of Posts and Telecommunications[3]

{jinkai, yangying.angulia, gaomingzhi}@bigo.sg sandykwokcs@stu.pku.edu.cn
{weizeqiang, xiuzhuang.zhou}@bupt.edu.cn

Abstract

Light Field (LF) cameras capture rich information in 4D LF images by recording both intensity and angular directions, making it crucial to learn the inherent spatial-angular correlation in low-resolution (LR) images for superior results. Despite impressive progress made by several CNN-based deep methods and pioneering Transformer-based methods for LF image super resolution (SR), most of them fail to fully leverage the LF spatial-angular correlation and tend to perform poorly in scenes with varying disparities. In this paper, we propose a hybrid method called DistgEPIT that implements an enhanced disparity learning mechanism with both convolution-based and transformer-based modules. It enables the capture of angular correlation, refinement of adjacent disparities, and extraction of essential spatial features. Additionally, we introduce a Position-Sensitive Windowing (PSW) strategy to maintain consistency of disparity between the training and inference stages, which yields an average PSNR gain of 0.2 dB by replacing the traditional padding and windowing method. Extensive experiments with ablation studies demonstrate the effectiveness of our proposed method, which ranked 1st place in the NITRE2023 LF image SR challenge. The code is available at https://github.com/OpenMeow/NTIRE23_LFSR_DistgEPIT.

1. Introduction

Light Field cameras [2] capture light rays from a scene in multiple directions, resulting in a more realistic 4D representation than traditional 2D images. This technology has a wide range of applications, such as post-capture refocusing [41], [2], [60], 3D reconstruction [3], [53], depth estimation [59], [17], [20], [23], de-occlusion [40], [56] and virtual reality [9], [51]. However, due to sensor limitations, there is a trade-off between spatial and angular resolution, making high-resolution LF images essential for many applications. LF images suffer from low spatial resolution, necessitating LF image super-resolution (SR) to leverage additional angular information and produce high-resolution SAIs with more details. Several CNN-based and Transformer-based LF methods have been proposed, achieving impressive progress. However, most of them fail to fully leverage the LF spatial-angular correlation and encounter performance bottlenecks in scenes with varying disparities.

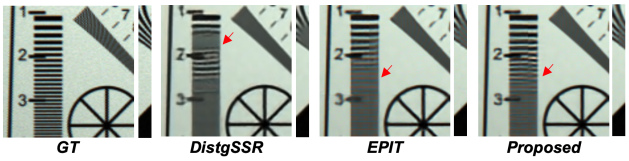

Figure 1. Super-resolved sample visualization. The red arrow highlights the major difference among DistgSSR [38], EPIT [27] and DistgEPIT (proposed), where our method produces clearer details and sharper textures.

In general, neighboring regions of the same pixel position in different sub-aperture images exhibit similar structural relationships. Therefore, a large number of CNN-based methods have been used, benefiting from their excellent local representation capability. For instance, [50] proposed LFCNN as the first CNN-based LF image SR method in the deep learning era. Subsequently, different works have further improved the SR performance through various CNN architectures. For example, [52] adopted a combined CNN to enhance Epipolar Plane Images (EPIs), [46] fused a set of sheared EPIs with CNN, [6] introduced a spatial-angular

*Equal contribution.

†Corresponding author. This work was supported by the National Natural Science Foundation of China under grants 61972046.

separable convolution, and [55] used a residual CNN architecture. Recently, learning angular information has received more attention in performance improvement. For instance, [38] proposed a generic mechanism to disentangle 4D LF data into a subspace and fully addressed the varying disparities caused by the angular dimension. In addition, [36] integrated such a disentangling mechanism and extended it to multiple degradations.

However, the position over boundaries presents a large disparity in light field images, which requires the method to aggregate remote features among different SAIs. Therefore, transformer-based methods are proposed to effectively model long-range information. Referring to prior works, LFT [26] successfully adapted Transformer into LF image processing, [27] proposed EPIT to learn better spatial-angular correlation through re-organized EPIs. Regarding the different advantages existed in both CNN-based and transformer-based network, we managed to incorporate the two kinds of architectures to address both adjacent and long-ranged disparities.

In addition, we found that the general padding method used in most SISR task can destroy the disparity relationship and produce worse predictions, since the sub-aperture views from LF cameras have strict optical disparity constraints. Therefore, we propose a Position-Sensitive Windowing (PSW) operation that maintains disparity structural consistency in the SAI subspace during windowing, without additional padding and disparity variations.

Our main contribution can be summarized as:

- We combine a convolution-based local correlation module and a transformer-based non-local correlation module to model adjacent and long-range disparity variations, resulting in finer details and textures;
- We introduce the Position-Sensitive Windowing (PSW) operation to address the issue of the disparity structure breaking caused by general padding methods, achieving a gain of 0.2 dB in average PSNR;
- Our method obtains state-of-the-art results with an average PSNR of **30.66 dB** across real and synthetic datasets, and ranked first place in NTIRE 2023 Light Field Super Resolution Challenge [37].

2. Related Work

LF image SR is a long term research topic, the earlier works followed different formulation theories yet traditional paradigm to resolve the task: [2, 44] conducted variational analysis to reconstruct super-resolved texture, [1, 12, 31] chose to focus on patches while [31] adopted a Gaussian mixture model (GMM), [12] learned linear projections from patch-volumes, [1] used BM3D and extended it into the proposed LFBM5D as a patches filter, [14] proposed a graph-based strategy. All above earlier works advanced the LF image SR research but their performance were also limited due to the hand-crafted image priors were incapable of extracting spatial features effectively.

Therefore CNN-based LR image SR approaches became dominant in deep learning era. [50], [15] became the earliest works that adopted CNNs. Then [52] used a combined network where a SISR CNN designed for SAIs spatial resolution enhancing and another designed for Epipolar Plane Images (EPIs) learning. [55] used residual convolutional neural networks to further improve spatial feature extraction and [13] used to restore LF low-rank prior. [18] continued spatial SR learning via an All-to-One network including combinatorial geometry embedding and structural consistency regularization module for parallax preservation. [35] introduced novel bidirectional recurrent CNN, [6] utilized spatial-angular separable (SAS) convolutions as approximating 4D convolution and [8] used CNN to aggregate warped SAIs, [42] proposed deformable convolution network LF-DFnet. More recently, [47] and [7] continued improving this task with zero-shot learning, [36] expanded the capability of handling various degradations among LF images in their proposed LF-DANet, [38] proposed DistgSSR network and introduced a generic yet effective disentangling mechanism through Spatial/Angular/Epipolar Plane Feature Extractors.

Recently, due to the long-range modeling ability, transformer-based model is widely applied on varied vision tasks. Vision Transformers such as ViT [10] for image classification, DETR [4] for object detection, SETR [58] for semantic segmentation, had achieved impressive performance in basic computer vision tasks. For low-level vision task, transformer-based methods also achieved an excellent performance [5,19,24,25,30,43,48]. Benefit from the long-range modeling ability for large disparity variations, Liang etc [26] pointed out the deficiency of CNN and designed two Transformers, one for incorporating angular information among different views and another for capturing spatial information. Most recent EPIT [27] expanded deeper into the large disparity variations existed in LF image, modeled the long-range dependencies over EPIs.

3. Method

3.1. Network Design

Inspired by prior works [27, 38], we propose the DistgEPIT network, which incorporates a correlation module to learn correspondence relationships between sub-views in both horizontal and vertical directions, a local correlation module to model precise spatial-angular information, plus a hierarchical fusion strategy to optimize performance. The overall architecture is presented in Figure 2.

In general, a light field image is expressed as a 4D tensor $\mathcal{I}(u, v, h, w) \in \mathbb{R}^{U \times V \times H \times W}$, where H and W denote

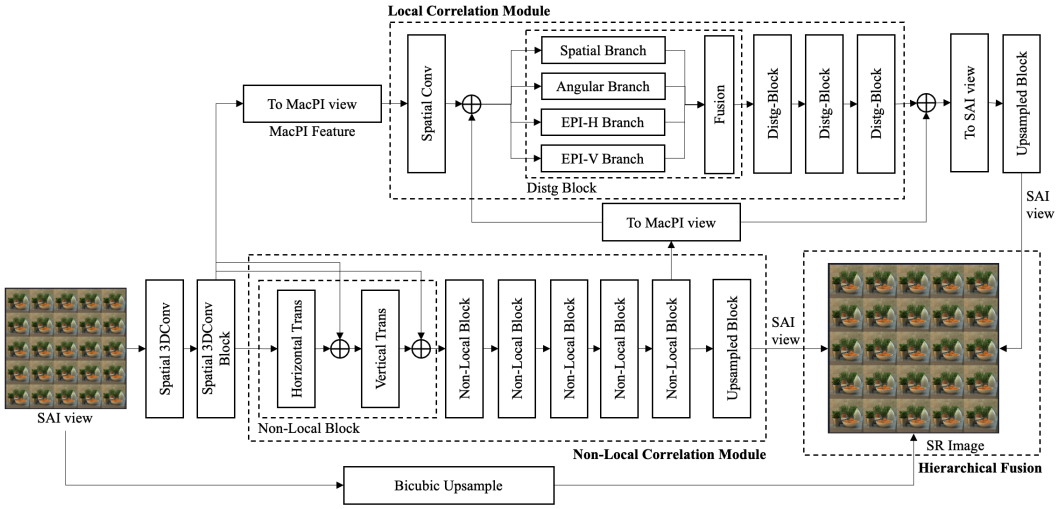

Figure 2. **Architecture for our proposed DistgEPIT.** Given SAIs as inputs, we adopt Bicubic Upsampling to restore the low-frequency content of the original images. For feature acquisition, we design a Transformer block-based Correlation Module to capture long-range disparity information. Meanwhile, we transform SAIs into Macro-Pixel Image (MacPI) views and leverage multiple CNN feature extractors to further refine high-frequency details and textures. During the final stage, we combine upsampled SAIs through Hierarchical Fusion to obtain ultimate super-resolved SAI results.

the spatial dimensions, and U and V denote the horizontal and vertical angular dimensions. Given a LR SAI array $\mathcal{I}_{LR}^{SAI} \in \mathbb{R}^{UH \times VW}$, the network finally outputs a HR SAI array $\mathcal{I}_{HR}^{SAI} \in \mathbb{R}^{\alpha AH \times \alpha AW}$, where α denotes an upsampled factor.

Non-Local Correlation Learning. In scenes with large disparity variations, unsatisfactory correlation learning can leads to a significant performance gap. To overcome this, we leverage the long-range modeling ability of the Transformer structure and adopt Epipolar Plane Images (EPIs) views in the first-stage feature learning, as previously proposed by methods like [27]. In EPI views, features are represented using oriented lines, which can encode disparity values effectively.

Initially, the low-resolution 4D light field (LF) image is upsampled using bicubic interpolation to a size of $\alpha H \times \alpha W$. Meanwhile, it is converted to $\mathcal{F}_{in} \in \mathbb{R}^{1 \times UV \times H \times W}$ format and passed through a series of $1 \times 3 \times 3$ spatial convolutional layers with Leaky ReLU activation to obtain a high-dimensional feature representation as $\mathcal{F}_{init} \in \mathbb{R}^{C \times UV \times H \times W}$. The number of channels in all convolutional layers is set to C for stable learning in restoration tasks.

Then the initial feature $\mathcal{F}_{init}$ is fed into a module to capture long-range information from the epipolar line. We follow the approach of EPIT [27] and define a series of cascading blocks as follows:

$$\mathcal{F}_{trans}^{EPI} = M_{trans}^{C,B_t,U,V}(\mathcal{F}_{init}^{EPI}) \tag{1}$$

Where, C denotes the number of channels, B_t denotes the number of blocks, and U and V denote the angular resolution. Each cascading block consists of ordered horizontal and vertical feature extractors, requiring the input feature in the form of $\mathcal{F}_h^{EPI} \in \mathbb{R}^{C \times UH \times V \times H}$ or $\mathcal{F}_v^{EPI} \in \mathbb{R}^{C \times VW \times U \times H}$ in EPI format.

Additionally, the initial feature is converted to the Macro-Pixel Image (MacPI) view and passed through a local correlation module built with convolutional blocks to learn strict correspondences. It is worth noting that placing the transformer-based blocks at the initial position is crucial, as CNN feature extractors based on local receptive field learning can easily destroy long-range information. In our experiments, we found that the paradigm of convolution block before transformer-based block is unable to converge. Further discussion in Section 4.3.

Local Correlation Learning. Although EPI views are suitable for learning long-range information using transformer-based blocks, they are not effective in modeling compact neighboring features and incorporating spatial context prior. To overcome this limitation, we use disentangling blocks proposed in [38] to extract fine-grained features with domain-specific convolutions.

After the correlation module aggregates the features with distinct information of large disparity variations, they would have been converted into a MacPI view with $\mathcal{F}^{MacPI} \in \mathbb{R}^{C \times UH \times VW}$, where the spatial and angular features are evenly mixed, making which more effective to convolve structured details. Then MacPI features would pass through a series of disentangling blocks. The convolution-based re-

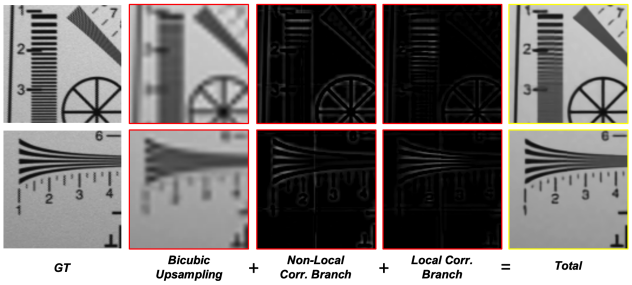

Figure 3. Feature fusion illustration. In our processing flow, Bicubic Upsampling simply provides rough super-resolved content, the use of correlation branch obtains clear boundaries and local correlation branch captures the high-frequency details.

finement operation follows:

$$\mathcal{F}_{conv}^{MacPI} = M_{conv}^{C,G_c,B_c,U,V}(\mathcal{F}_{init}^{MacPI}, \mathcal{F}_{trans}^{MacPI}) \quad (2)$$

Here, C is the number of channels, G_c is the number of groups, B_c is the number of blocks, and U and V are the angular resolutions. Each block comprises four parallel branches in the spatial, angular, horizontal, and vertical EPI domain-specific convolutions. Assuming an angular resolution $U = V = A$, the spatial convolution has a kernel size of 3×3 with dilation of A, the angular convolution has a kernel size of $A \times A$ with a stride of A, and the horizontal and vertical EPI convolutions have a kernel size of $1 \times A^2$ with a stride of A, using identical weight parameters.

Hierarchical Fusion. Due to the feature modeling discrepancy between transformer-based extractors and convolution-based extractors, we use three super-resolved SAI formatted images to combine the information from above two types of extractors. As shown in Figure 3, it allows the optimizer to guide the transformer-based extractor on modeling long-range information and convolution-based extractor on modeling compact local spatial information independently. The fused output is given as:

$$\mathcal{I}_{out}^{SAI} = \alpha U_b(\mathcal{I}_{in}^{SAI}) + \beta U_t(\mathcal{F}_{trans}^{EPI}) + \gamma U_c(\mathcal{F}_{conv}^{MacPI}) \quad (3)$$

where Ub represents Bicubic Upsampling of the original low-resolution image, U_t applies upsampling to the feature from the transformer-based blocks using a convolution layer with a 3×3 kernel size, and U_c applies upsampling to the feature from the convolution-based disentangling blocks using a convolution layer with a 1×1 kernel size. The coefficients α, β, and γ are set to 1.0, 0.5, and 0.5 respectively.

3.2. Position-Sensitive Windowing Operation

While the proposed method using Transformer-based blocks and convolution-based blocks can effectively learn disparity features, the post-processing method based on center padding breaks the disparity structural correlation in the SAI subspaces. For SISR task, center padding and windowing methods are typically used for seamlessly stitching results obtained by sliding windows, then generate the final super-resolved image. Center padding ensures the correctness of edge pixels and avoids edge artifacts caused by sliding windows. Windowing methods can also reduce the extra computation required for each block and corresponding memory usage, thereby improve computational efficiency and lead to faster generation of super-resolved images.

However, the sub-aperture views captured by LF cameras have strict optical disparity constraints. Each subspace in the light field image exhibits significant spatial-angular correlation, while the disparity values gradually decrease from the outermost layer to the center. Due to introduction of artifact padding values with unfaithful disparity structure, center padding clearly destroys the disparity relationship in subspace, which makes position-sensitive learning-based networks produce worse predictions. Therefore, we propose a Position-Sensitive Windowing (PSW) operation that ensures the disparity structural consistency in the SAI subspace maintained during windowing without additional computation.

Assuming the SAI image forms $\mathcal{I} \in \mathbb{R}^{C \times U \times V \times H \times W}$, the stride and window size of the PSW operation are set to S and K, respectively. Without introducing any padding operations, the operation adopts a sliding window approach to crop the block in an overlapping manner, and for border values, it backtracks to fill in the entire block. The total number of blocks can be formulated as:

$$N = \lfloor \frac{H + S - 1}{S} \rfloor \cdot \lfloor \frac{W + S - 1}{S} \rfloor \quad (4)$$

The proposed PSW operation does not introduce extra computational cost. Detailed padding fashions are discussed in Section 4.3.

4. Experiments

In this section, we firstly describe our experimental details, then we elaborate our comprehensive ablation studies and thorough performance comparison with other methods.

4.1. Datasets and Implementation Details

We used the five public LF datasets: EPFL [32], HCInew [16], HCIold [45], INRIA [22], and STFgantry [33], following the same training and testing partition as in [42]. All datasets have the same angular resolution of 9×9. During training, we selected a 5×5 SAI view and extracted patches with a stride of 32. Then we performed bicubic downsampling to generate low-resolution light field patches as input samples. Additionally, data augmentation techniques such as horizontal flipping, vertical flipping, and 90-degree rotation were applied.

Table 1. Overall PSNR/SSIM metrics comparison among the other prestigious approaches for $4\times$ SR. We have obtained state-of-the-art results among all 5 datasets. The best averaged results are achieved by our DistgEPIT†-TTA method(highlighted in bold fonts).

Methods	EPFL	HCInew	HCIold	INRIA	STFganry	Average
Bicubic	25.14 / 0.8324	27.61 / 0.8517	32.42 / 0.9344	26.82 / 0.8867	25.93 / 0.8452	27.58 / 0.8701
VDSR [21]	27.25 / 0.8777	29.31 / 0.8823	34.81 / 0.9515	29.19 / 0.9204	28.51 / 0.9009	29.81 / 0.9066
EDSR [28]	27.84 / 0.8854	29.60 / 0.8869	35.18 / 0.9536	29.66 / 0.9257	28.70 / 0.9072	30.20 / 0.9118
RCAN [57]	27.88 / 0.8863	29.63 / 0.8886	35.20 / 0.9548	29.76 / 0.9276	28.90 / 0.9131	30.27 / 0.9141
resLF [55]	28.27 / 0.9035	30.73 / 0.9107	36.71 / 0.9682	30.34 / 0.9412	30.19 / 0.9372	31.25 / 0.9322
LFSSR [49]	28.27 / 0.9118	30.72 / 0.9145	36.70 / 0.9696	30.31 / 0.9467	30.15 / 0.9426	31.23 / 0.9370
LF-ATO [18]	28.52 / 0.9115	30.88 / 0.9135	37.00 / 0.9699	30.71 / 0.9484	30.61 / 0.9430	31.54 / 0.9373
LF-InterNet [39]	28.67 / 0.9162	30.98 / 0.9161	37.11 / 0.9716	30.61 / 0.9491	30.53 / 0.9409	31.58 / 0.9388
LF-DFnet [42]	28.77 / 0.9165	31.23 / 0.9196	37.32 / 0.9718	30.83 / 0.9503	31.15 / 0.9494	31.86 / 0.9415
MEG-Net [54]	28.74 / 0.9160	31.10 / 0.9177	37.27 / 0.9716	30.66 / 0.9490	30.77 / 0.9453	31.71 / 0.9399
LF-IINet [29]	29.11 / 0.9188	31.36 / 0.9208	37.62 / 0.9734	31.08 / 0.9515	31.21 / 0.9502	32.08 / 0.9429
DPT [34]	28.93 / 0.9170	31.19 / 0.9188	37.39 / 0.9721	30.96 / 0.9503	31.14 / 0.9488	31.92 / 0.9414
LFT [26]	29.25 / 0.9210	31.46 / 0.9218	37.63 / 0.9735	31.20 / 0.9524	31.86 / 0.9548	32.28 / 0.9447
DistgSSR [38]	28.99 / 0.9195	31.38 / 0.9217	37.56 / 0.9732	30.99 / 0.9519	31.65 / 0.9535	32.11 / 0.9440
EPIT [27]	29.34 / 0.9197	31.51 / 0.9231	37.68 / 0.9737	31.27 / 0.9526	32.18 / 0.9571	32.40 / 0.9452
DistgEPIT	30.09 / 0.9224	31.61 / 0.9252	37.96 / 0.9742	32.35 / 0.9535	32.45 / 0.9589	32.90 / 0.9468
DistgEPIT†	30.17 / 0.9232	31.71 / 0.9263	38.03 / 0.9744	32.39 / 0.9535	32.74 / 0.9604	33.01 / 0.9476
DistgEPIT†-TTA	**30.41 / 0.9260**	**31.91 / 0.9283**	**38.24 / 0.9753**	**32.60 / 0.9551**	**33.06 / 0.9626**	**33.25 / 0.9495**

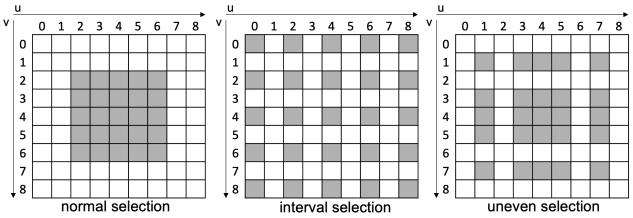

Figure 4. Sampling Strategy illustration. Central Selection(CS) is most commonly used by selecting center part, while our additional Interval Selection(IS) and Uneven Selection (US) methods sample those near boundaries to enlarge disparities.

To further enhance large disparity learning, we introduced two additional sampling methods, as shown in Figure 4. The Central Selection (CS) applied the same as most previous works. The Interval Selection (IS) and the Uneven Selection (US) tended to sample Sub-Aperture Images (SAI) closer to the borders, resulting in larger disparity variations than those selecting from center. Noticeably, in the rest of this paper, we use the symbol † to denote the proposed method trained with a combination of CS, IS, and US, whose result brought a three-fold increase in training time due to the enlarged size of the dataset.

For optimization, we adopted the L1 loss function and Adam optimizer with a learning rate of 2e-4, β_1 = 0.9, and β_2 = 0.999. Then we distributed four samples equally on four NVIDIA RTX 2080Ti GPUs to train. At 80 epochs, we decayed the learning rate to half of the former value and stop training at 100 epochs.

As conventional usage in related works [28, 29, 42, 57] , we use the PSNR and SSIM computed only on the Y channel of images as quantitative metrics for performance evaluation. To compute the metric scores for a dataset containing M scenes, we firstly compute the average score of each scene by separately averaging the scores over all SAIs. Then metric score for the dataset is determined by averaging the scores over the M scenes.

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

We compared DistgEPIT to several state-of-the-art methods, including three SISR methods: VDSR [21], EDSR [28], and RCAN [57] and other eleven recent LF image SR methods: resLF [55], LFSSR [49], LF-ATO [18], LF-InterNet [39], LF-DFnet [42], MEG-Net [54], LF-IINet [29], DPT [34], LFT [26], DistgSSR [38], and EPIT [27]. Additionally, a Bicubic Upsampling method is introduced as baseline. DistgEPIT† is trained on an extended selection strategy, and DistgEPIT†-TTA employs the test-time-augmentation technique(TTA) by using seven different affine transformations.

Quantitative Results. As shown in Table 1, regarding quantitative performance, the proposed DistgEPIT achieves state-of-the-art 32.90 PSNR and 0.9468 SSIM scores for the $4\times$ LFSR task. Compared to the second best performing method EPIT, our DistgEPIT gains additional 0.75 dB, 0.1 dB, 0.28 dB, 1.08 dB, and 0.27 dB over five datasets respectively. Benefits from the non-correlation and local correlation module, plus Position-Sensitive Windowing (PSW) operation, our average PSNR achieves a significant 0.50dB improvement.

Qualitative Results. As shown in Figure 5, for qualitative results, the proposed DistgEPIT demonstrates a well-reconstruction ability to generate faithful details and sharp

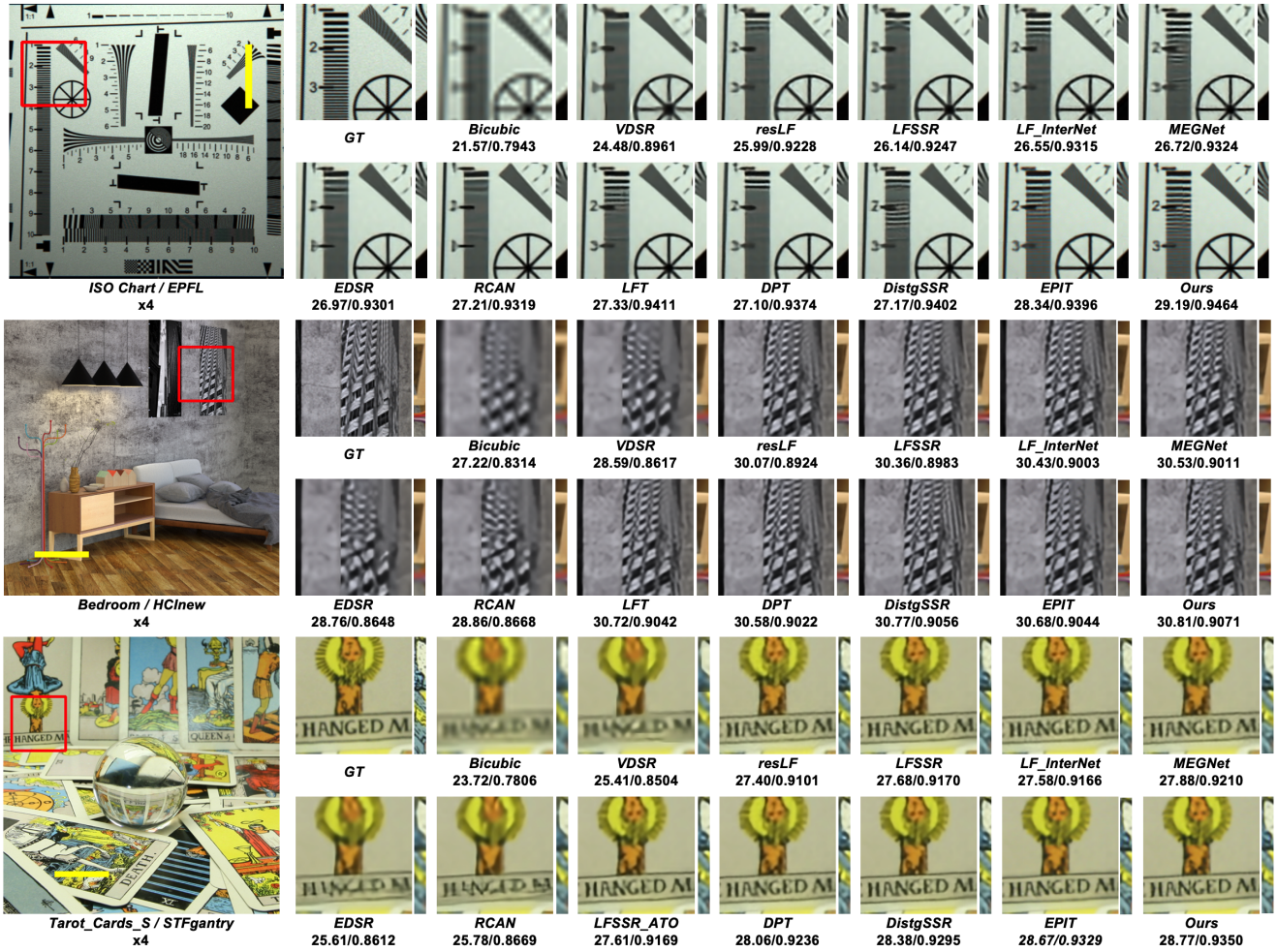

Figure 5. Qualitative results for 4×SR. The super-resolved center view images are presented for detailed texture comparison. The corresponding PSNR/SSIM scores of different methods on the presented scenes are also reported below.

structures with hardly any artifacts. For SISR methods, VDSR, EDSR and RCAN tend to generate flat images with little detail. For recent LFSR methods, the proposed DistgEPIT could distinguish more dense detail information (e.g., clearer strip lines of ISO Chart).

4.3. Ablation Study

In this section, we further consolidate the effectiveness of every key component within our DistgEPIT via ablation study experimental results.

Particularly, our method combines a transformer-based correlation module and a convolution-based local correlation module with channel size of 128, and a single sample will occupy the entire NVIDIA RTX 2080Ti GPU memory space. To ensure fair comparison, we increased the channel number of DistgSSR from 64 to 160 to match similar parameter volume and computational complexity of DistgEPIT. Moreover, the channel size of EPIT was modified to 180 to fit in a single GPU memory block due to GPU memory constraints.

Model Scale. The increase in network channel numbers significantly improves the accuracy of the model on the LFSR dataset. Compared to their baseline models, DistgSSR-C160 and EPIT-C180 improve the PSNR metric by 0.29 dB and 0.18 dB respectively. Despite of the similar parameter and computation cost, DistgEPIT outperforms DistgSSR-C160 with a 0.13 dB PSNR improvement, demonstrating the efficiency of the proposed archietecture.

Convolution-First. An intuitive question can arises: In what order shall we combine the two modules? We designed DistgEPIT-Inv with the same computational complexity as DistgEPIT's comparison, utilizing sub-modules similar to the DistgEPIT structure. In DistgEPIT-Inv, the input image is firstly processed by the convolution-based local correlation module to extract features, which are then fed into the transformer-based correlation module to be refined.

Table 2. A comparison of PSNR/SSIM metrics among different model volumes and model orders for 4x super-resolution, with the best results highlighted in bold. Note that all methods are retrained with same settings of DistgEPIT with PSW strategy.

Methods	#Param.	FLOPs	EPFL	HCInew	HCIold	INRIA	STFganry	Average
DistgSSR-C64	3.58 M	65.41 G	29.78	31.41	37.69	32.00	31.52	32.48
DistgSSR-C128	14.33 M	261.25 G	30.01	31.52	37.77	32.25	31.81	32.67
DistgSSR-C160	22.37 M	408.09 G	30.04	**31.63**	37.94	32.26	32.01	32.77
EPIT-C64	1.13 M	33.36 G	29.81	31.45	37.73	32.05	32.07	32.62
EPIT-C128	4.55 M	132.20 G	29.97	31.57	37.83	32.21	32.25	32.77
EPIT-C180	9.01 M	260.84 G	30.02	**31.63**	37.86	32.20	32.30	32.80
DistgEPIT-Inv	19.02 M	397.23 G	29.95	31.55	37.90	32.19	32.31	32.78
DistgEPIT	19.02 M	397.20 G	**30.09**	31.61	**37.96**	**32.35**	**32.45**	**32.90**

Table 3. Quantitative PSNR comparison among different padding strategies shows that PSW (Position-Sensitive Windowing) brings performance gain across all three methods. Note that all methods are retrained with same settings of DistgEPIT with PSW strategy. Center offset pads both upper left and bottom right area, zero offset pads bottom right area, while PSW does not pad any areas.

Method	Offset	Padding	EPFL	HCInew	HCIold	INRIA	STFganry	Average
DistgSSR-C64	Center	Mirror	27.74	31.37	37.51	29.56	31.07	31.45
		Replicate	28.45	31.38	37.44	30.41	31.16	31.77
	Zero	Mirror	29.57	31.41	37.67	31.78	31.43	32.37
		Replicate	29.64	31.40	37.69	31.87	31.45	32.41
	-	**PSW**	**29.78**	**31.41**	**37.69**	**32.00**	**31.52**	**32.48**
EPIT-C64	Center	Mirror	27.67	31.37	37.28	29.49	31.42	31.45
		Replicate	28.32	31.40	37.30	30.32	31.68	31.81
	Zero	Mirror	29.50	31.44	37.69	31.80	31.95	32.48
		Replicate	29.60	31.44	37.69	31.90	31.98	32.52
	-	**PSW**	**29.81**	**31.45**	**37.73**	**32.05**	**32.07**	**32.62**
DistgEPIT	Center	Mirror	27.97	31.55	37.71	29.79	31.81	31.77
		Replicate	28.59	31.56	37.63	30.62	31.97	32.07
	Zero	Mirror	29.87	31.61	37.93	32.13	32.34	32.78
		Replicate	29.88	31.60	37.93	32.20	32.36	32.79
	-	**PSW**	**30.09**	**31.61**	**37.96**	**32.35**	**32.45**	**32.90**

Besides DistgEPIT-Inv and DistgEPIT employ identical hierarchical fusion strategy and loss function. As revealed in Table 2, after the module order swapping, DistgEPIT-Inv's performance drops by an average of 0.12 dB compared to DistgEPIT and typically drops by 0.14 dB on the STFganry dataset with larger disparities. Which suggests that the convolution-based module may lose long-range disparity information in the initial stage, making it difficult to refine the features in the subsequent correlation learning.

Windowing Operation. The effectiveness of different padding methods with zero offset is presented in Table 3. Zero offset padding methods outperform center padding methods, primarily due to the preservation of the disparity relationship in the top-left corner. Among the zero offset methods, replicate padding is superior to mirror padding as it preserves the increasing disparity structure from the center to the edge by copying the last boundary element during padding. However, the disparity structure introduced by padding is still an artificially designed one. Hence, the proposed Position-Sensitive Windowing (PSW) strategy strictly enforces the natural disparity structure on the last row or column of the partitioned window, leading to better performance than replicate padding of zero offset by 0.07 dB, 0.10 dB, and 0.11 dB on the DistgSSR, EPIT, and DistgEPIT models respectively. It is noteworthy that the PSW strategy is particularly effective in transformer-based networks as it relies on the strict disparity structure relationship among SAIs to learn long-range information. Any unrealistically introduced disparity information may result in querying incorrect features, which can introduce noise to the feature representation in subsequent modules.

SAIs Selection. Table 4 demonstrates that the proposed DistgEPIT method achieves a significant improvement of 0.20 dB on the STFganry dataset with large disparities by incorporating the Interval Selection (IS) strategy, with an average improvement of 0.09 dB. Additionally, the Uneven

Table 4. Quantitative results show that each additional sampling strategy can lead to further performance increment. (CS: Central Selection, IS: Interval Selection, US: Uneven Selection).

CS	IS	US	EPFL	HCInew	HCIold	INRIA	STFganry	Average
√			30.09	31.61	37.96	32.35	32.46	32.90
√	√		30.16 (+0.07)	31.70 (+0.09)	**38.05 (+0.09)**	32.36 (+0.01)	32.66 (+0.20)	32.99 (+0.09)
√	√	√	**30.17 (+0.08)**	**31.71 (+0.10)**	38.03 (+0.07)	**32.39 (+0.04)**	**32.74 (+0.28)**	**33.01 (+0.11)**

Table 5. Our team achieved 1st place on the leader board (last three rows) in the NTIRE-2023 test dataset, with quantitative results of 30.6640 dB PSNR (average) and 0.9314 SSIM (average). For a single model comparison, the proposed DistgEPIT achieves an average PSNR of 30.275 dB and an average SSIM of 0.9273 with using central selection only.

Methods	#Params.	Lytro	Synthetic	Average
Bicubic	–	25.109 / 0.8404	26.461 / 0.8352	25.785 / 0.8378
VDSR [21]	0.665 M	27.052 / 0.8888	27.936 / 0.8703	27.494 / 0.8795
EDSR [28]	38.89 M	27.540 / 0.8981	28.206 / 0.8757	27.873 / 0.8869
RCAN [57]	15.36 M	27.606 / 0.9001	28.308 / 0.8773	27.957 / 0.8887
resLF [55]	8.646 M	28.657 / 0.9260	29.245 / 0.8968	28.951 / 0.9114
LFSSR [49]	1.774 M	29.029 / 0.9337	29.399 / 0.9008	29.214 / 0.9173
LF-ATO [18]	1.364 M	29.087 / 0.9354	29.401 / 0.9012	29.244 / 0.9183
LF_InterNet [39]	5.483 M	29.233 / 0.9369	29.446 / 0.9028	29.340 / 0.9198
MEG-Net [54]	1.775 M	29.203 / 0.9369	29.539 / 0.9036	29.371 / 0.9203
LF-IINet [29]	4.886 M	29.487 / 0.9403	29.786 / 0.9071	29.636 / 0.9237
DPT [34]	3.778 M	29.360 / 0.9388	29.771 / 0.9064	29.566 / 0.9226
LFT [26]	1.163 M	29.657 / 0.9420	29.881 / 0.9084	29.769 / 0.9252
DistgSSR [38]	3.582 M	29.389 / 0.9403	29.884 / 0.9084	29.637 / 0.9244
LFSSR_SAV [49]	1.543 M	29.713 / 0.9425	29.850 / 0.9075	29.782 / 0.9250
EPIT [27]	1.470 M	29.718 / 0.9420	30.030 / 0.9097	29.874 / 0.9259
HLFSR-SSR [11]	13.87 M	29.714 / 0.9429	29.945 / 0.9097	29.830 / 0.9263
DistgEPIT	19.02 M	30.408 / 0.9436	30.141 / 0.9109	30.275 / 0.9273
DistgEPIT†	19.02 M	30.485 / 0.9443	30.299 / 0.9127	30.392 / 0.9285
DistgEPIT†-TTA	19.02 M	30.746 / 0.9468	30.460 / 0.9146	30.603 / 0.9307
OpenMeow	/	30.82 / 0.9475	**30.51** / 0.9152	**30.66** / 0.9314
DMLab	/	**30.92** / 0.9489	30.35 / 0.9146	30.64 / 0.9318
VIDAR	/	30.67 / **0.9491**	30.45 / **0.9154**	30.56 / **0.9323**

Selection (US) strategy further enhances the performance on STFganry by 0.08 dB and the average improvement by 0.02 dB. The minor gain observed from the US strategy indicates an underlying issue related to the strict optical disparity constraints causing damage.

4.4. NTIRE 2023 LFSR Challenge Results

The NTIRE 2023 LFSR challenge develop a new dataset, named NTIRE-2023 [37], where the 16 synthetic LFs and 16 real-world LFs captured by Lytro camera for test subset. During the challenge, all participants were strictly prohibited from using any external model or data, including pre-trained backbones and optical flow networks. For the final results reporting, we used the average ensemble method to combine the outputs generated by the DistgEPIT with different configurations and the DistgSSR with different scales. As shown in Table 5, proposed method ranked the 1st place with 30.6640 dB PSNR on the LFSR test dataset.

5. Conclusion and Future Work

In this paper, we investigated the task of Light Field image Super Resolution (LF image SR), in which we addressed the issue of large disparities not being fully utilized during the super-resolving process. To that end, we proposed a CNN-Transformer hybrid network called DistgEPIT. The proposed network could learn better long-range angular correlation with the help of transformer-based correlation module, while maintaining robust spatial features and adjacent correlation via convolution-based local correlation module. Additionally, we introduced a novel Position-Sensitive windowing (PSW) operation to maintain the disparity correspondence. Our proposed method achieved leading performance with a PSNR of **30.6640 dB**, and it won the **1st place** in the NTIRE 2023 Light Field Super Resolution contest track.

In our future work, we will explore the disparity problem and further improve the performance of the current hybrid framework.

References

[1] Martin Alain and Aljosa Smolic. Light field super-resolution via lfbm5d sparse coding. 10 2018. 2

[2] Tom Bishop and Paolo Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34:972 – 986, 06 2012. 1, 2

[3] Zewei Cai, Xiaoli Liu, Xiang Peng, and Bruce Gao. Ray calibration and phase mapping for structured-light-field 3d reconstruction. *Optics Express*, 26:7598, 03 2018. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. *End-to-End Object Detection with Transformers*, pages 213–229. 11 2020. 2

[5] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer, 05 2022. 2

[6] Zhen Cheng, Yutong Liu, and Zhiwei Xiong. Spatial-angular versatile convolution for light field reconstruction. *IEEE Transactions on Computational Imaging*, 8:1131–1144, 2022. 1, 2

[7] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10005–10014, 2021. 2

[8] Zhen Cheng, Zhiwei Xiong, and Dong Liu. Light field super-resolution by jointly exploiting internal and external similarities. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2604–2616, 2020. 2

[9] 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*, 40:1–12, 12 2021. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 10 2020. 2

[11] V. V. Duong, T. H. Nguyen, J. Yim, and B. Jeon. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Trans. Compuational Imaging*, 2023. 8

[12] Reuben Farrugia, Christian Galea, and Christine Guillemot. Super resolution of light field images using linear subspace projection of patch-volumes. *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1, 08 2017. 2

[13] Reuben A. Farrugia and Christine Guillemot. Light field super-resolution using a low-rank prior and deep convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1162–1175, 2020. 2

[14] Vahid Khorasani Ghassab and Nizar Bouguila. Light field super-resolution using edge-preserved graph-based regularization. *IEEE Transactions on Multimedia*, 22(6):1447–1457, 2020. 2

[15] M. Shahzeb Khan Gul and Bahadir K. Gunturk. Spatial and angular resolution enhancement of light fields using convolutional neural networks. *IEEE Transactions on Image Processing*, 27(5):2146–2159, 2018. 2

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

[17] 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. 1

[18] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. pages 2257–2266, 06 2020. 2, 5, 8

[19] Kai Jin, Zeqiang Wei, Angulia Yang, Sha Guo, Mingzhi Gao, Xiuzhuang Zhou, and Guodong Guo. Swinipassr: Swin transformer based parallax attention network for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 920–929, June 2022. 2

[20] N. Khan, M. H. Kim, and J. Tompkin. Differentiable diffusion for dense depth estimation from multi-view images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8908–8917, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. 1

[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 5, 8

[22] Mikael Le Pendu, Xiaoran Jiang, and Christine Guillemot. Light field inpainting propagation via low rank matrix completion. *IEEE Transactions on Image Processing*, 27(4):1981–1993, 2018. 4

[23] T. Leistner, R. Mackowiak, L. Ardizzone, U. Kuthe, and C. Rother. Towards multimodal depth estimation from light fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12951, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 1

[24] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. *CoRR*, abs/2201.12288, 2022. 2

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer, 2021. 2

[26] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. Light field image super-resolution with transformers. *IEEE Signal Processing Letters*, 29:563–567, 2022. 2, 5, 8

[27] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, Shilin Zhou, and Yulan Guo. Learning non-local spatial-angular correlation for light field image super-resolution. 02 2023. 1, 2, 3, 5, 8

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single

image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 5, 8

[29] Gaosheng Liu, Huanjing Yue, Jiamin Wu, and Jingyu Yang. Intra-inter view interaction network for light field image super-resolution. *IEEE Transactions on Multimedia*, 2021. 5, 8

[30] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng. Transformer for single image super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 456–465, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 2

[31] Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. pages 22–28, 06 2012. 2

[32] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, number CONF, 2016. 4

[33] V. Vaish and A. Adams. The (new) stanford light field archive. *Computer Graphics Laboratory*, 6(7), 2008. 4

[34] Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Detail-preserving transformer for light field image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2522–2530, 2022. 5, 8

[35] Yunlong Wang, Liu Fei, 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*, 27:1–1, 05 2018. 2

[36] Yingqian Wang, Zhengyu Liang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Learning a degradation-adaptive network for light field image super-resolution, 06 2022. 2

[37] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, Kai Jin, Zeqiang Wei, Angulia Yang, Sha Guo, Mingzhi Gao, Xiuzhuang Zhou, Vinh Van Duong, Thuc Nguyen Huu, Jonghoon Yim, Byeungwoo Jeon, Yutong Liu, Zhen Cheng, Zeyu Xiao, Ruikang Xu, Zhiwei Xiong, Gaosheng Liu, Manchang Jin, Huanjing Yue, Jingyu Yang, Chen Gao, Shuo Zhang, Song Chang, Youfang Lin, Wentao Chao, Xuechun Wang, Guanghui Wang, Fuqing Duan, Wang Xia, Yan Wang, Peiqi Xia, Shunzhou Wang, Yao Lu, Ruixuan Cong, Hao Sheng, Da Yang, Rongshan Chen, Sizhe Wang, Zhenglong Cui, Yilei Chen, Yongjie Lu, Dongjun Cai, Ping An, Ahmed Salem, Hatem Ibrahem, Bilel Yagoub, Hyun-Soo Kang, Zekai Zeng, and Heng Wu. Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 2, 8

[38] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 02 2022. 1, 2, 3, 5, 8

[39] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 290–308. Springer, 2020. 5, 8

[40] Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, and Yulan Guo. Deoccnet: Learning to see through foreground occlusions in light fields. pages 118–127, 03 2020. 1

[41] Yingqian Wang, Jungang Yang, Yulan Guo, Chao Xiao, and Wei An. Selective light field refocusing for camera arrays using bokeh rendering and superresolution. *IEEE Signal Processing Letters*, 26(1):204–208, 2019. 1

[42] Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo. Light field image super-resolution using deformable convolution (tip 2020). *IEEE Transactions on Image Processing*, 30, 11 2020. 2, 4, 5

[43] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17662–17672, 2022. 2

[44] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36, 08 2013. 2

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

[46] 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. 1

[47] Zhaolin Xiao, Yinhai Liu, Haiyan Jin, and Christine Guillemot. Zepi-net: Light field super resolution via internal cross-scale epipolar plane image zero-shot learning. *Neural Processing Letters*, 08 2022. 2

[48] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. pages 5790–5799, 06 2020. 2

[49] Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018. 5, 8

[50] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and Inso Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, PP:1–1, 02 2017. 1, 2

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

[52] Yan Yuan, Ziqi Cao, and Lijuan Su. Light-field image super-resolution using a combined deep cnn based on epi. *IEEE Signal Processing Letters*, PP:1–1, 07 2018. 1, 2

[53] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. pages 6505–6514, 10 2021. 1

[54] 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. 5, 8

[55] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. pages 11038–11047, 06 2019. 2, 5, 8

[56] Shuo Zhang, Zeqi Shen, and Youfang Lin. Removing foreground occlusions in light field using micro-lens dynamic filter. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1302–1308. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 1

[57] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 5, 8

[58] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. pages 6877–6886, 06 2021. 2

[59] Hao Zhu, Qing Wang, and Jingyi Yu. Occlusion-model guided antiocclusion depth estimation in light field. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):965–978, 2017. 1

[60] Hao Zhu, Qi Zhang, and Qing Wang. 4d light field superpixel and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6709–6717, 2017. 1