BigEPIT: Scaling EPIT for Light Field Image Super-Resolution*

Wentao Chao[1], Yiming Kan[1] Xuechun Wang[1], Fuqing Duan[1]†, Guanghui Wang[2]
[1]Beijing Normal University, [2] Toronto Metropolitan University
{chaowentao, kanyiming, wangxuechun}@mail.bnu.edu.cn, fqduan@bnu.edu.cn
wangcs@torontomu.ca

Abstract

Existing methods have been developed for light field (LF) image Super-Resolution (SR) and achieved continuously improved performance while suffering a significant performance drop when handling scenes with large disparity variations. EPIT [1] was proposed to mitigate the disparity issue through non-local spatial-angular correlation learning. However, EPIT has limitations due to the limited scale of existing LF datasets and the presence of imbalanced LF disparity, especially the scarcity of large disparity. To address this issue, we present a series of strategies to scale EPIT, called BigEPIT, including compound model scaling, augmented data resampling, and a high-precision test scheme. Specifically, the compound scaling method simultaneously scales the depth and width of the model to better improve the model capability. The augmented resampling method employs varying sampling intervals during training data generation, rather than solely relying on the central region view. This approach mitigates issues related to disparity imbalance and overfitting. The patch-based test scheme is popular because of its small GPU memory footprint. The traditional zero padding method and window partition will destroy the LF disparity structure and degrade the performance. Moreover, we find a positive correlation between the performance and the patchsize. Therefore, we advocate a high-precision test scheme i.e., a full-size or larger patchsize without zero padding for testing wherever the GPU memory permits, to achieve superior results. Extensive experiments demonstrate the effectiveness of our proposed method, which ranked 1st place in the NTIRE 2024 Light Field Image Super-Resolution Challenge.

1. Introduction

Light field (LF) images, captured in a single snapshot, not only record the spatial information of a scene but also contain abundant angular information. This has given rise to extensive practical applications, e.g., depth estimation [2–10], view synthesis [11, 12], 3D reconstruction [13], and virtual reality [14]. However, due to the inherent trade-off between spatial and angular resolution in LF images, it is challenging to obtain high spatial resolution with rich angular information or vice versa, limiting the practical applications of LF images. Therefore, numerous methods have been dedicated to enhancing the LF spatial or angular resolution, i.e., LF spatial or image super-resolution (SR) [15–39].

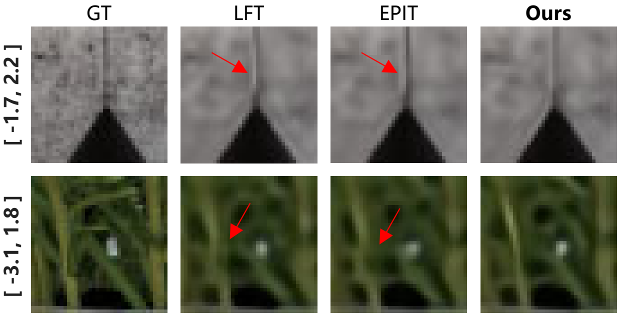

Figure 1. Visualization of 4× SR results of our method and state-of-the-art methods [1, 26] under different disparity values of LFs. Our method achieves superior performance and is robust to disparity variations.

In the early studies [15–20], researchers followed the traditional paradigm and proposed various models to formulate the problem. However, despite being able to capture the structure of LF, these models exhibited limited performance due to the inadequate representation capacity of handcrafted image priors. In recent years, convolutional neural networks (CNNs) have been effectively utilized in the field of LF image SR and made substantial advancements [21–35]. However, existing methods have achieved promising results on LF scenes with small baselines. However, their performance significantly deteriorates when handling scenes with large disparity variations. This may be attributed to the limited local receptive field of CNNs. Therefore, EPIT [1] was proposed to alleviate the disparity issue in LF images by uti-

*This work is supported by the National Key Research and Development Project Grant, Grant/Award Number: 2019YFC1521103.

†Corresponding author.

lizing horizontal and vertical epipolar Transformer to learn non-local spatial-angular correlations. However, EPIT still has room for improvement due to the limited scale of existing LF datasets and the imbalanced LF disparity, particularly the scarcity of large disparities. On the other hand, due to the large image size when testing, the patch-based test scheme is popular because of its small GPU memory footprint. The traditional zero padding method and window partition destroy the disparity structure of LFs and degrade the performance.

We aim to scale EPIT reasonably and further enhance the model's generalization ability to handle disparity variations of LF images. This paper presents a series of strategies to solve the above challenges, including compound model scaling, augmented data resampling, and a high-precision test scheme. Specifically, inspired by EfficientNet [40], we adapt the compound scaling method to simultaneously scale the depth and width of the model, called BigEPIT, which can better improve the model capability. The augmented resampling method specifies different sampling intervals when generating the training data rather than just using the view of the central region to alleviate disparity imbalance and overfitting problems. The augmented resampling method [41] addresses the issues of disparity imbalance and overfitting by generating training data with different sampling intervals, instead of solely relying on the central region view. Moreover, we have found that there is a positive correlation between the performance and the patchsize when the patch-based test scheme. We propose the adoption of a high-precision testing scheme, where a full-size or larger patch size is used without zero padding whenever the GPU memory allows to achieve superior results. Finally, our method achieves superior performance and is robust to disparity variation, as shown in Fig. 1.

In summary, the contributions of this work are as follows: (1) We propose a series of strategies that successfully scale the EPIT model to BigEPIT, which can better solve the disparity problem in LF image SR, including compound model scaling, augmented data resampling, and a high-precision test scheme. (2) Compared to existing state-of-the-art LF image SR methods, our method achieves superior performance i.e., average PSNR of **30.80 dB** on real and synthetic LF images, and won 1st place in the NTIRE 2024 Light Field Image Super-Resolution Challenge.

2. Related Work

2.1. Traditional Methods

Light Field Super-Resolution (LFSR) techniques primarily fall into two camps: those that rely on disparity, and those that leverage learning-based approaches to understand scene structures either directly or indirectly. Utilizing estimated structural information as a basis, numerous studies have concentrated on the challenge of accurately warping images from multiple viewpoints to access sub-pixel details, thereby enhancing spatial resolution. Notably, Wanner [42] utilized the structure tensor to deduce depth from Epipolar Plane Images (EPIs), employing this data to refine the resolution of view images through interpolation. Similarly, Mitra adopted a Gaussian Mixture Model to enhance LF patch resolution by leveraging disparity estimates. Other strategies [43] involve direct pixel warping from various views to restore image quality. Recently, Zhang et al. [44] introduced a method to correlate microlens and view images, using the texture-rich microlens photos for view image recovery. Rossi [20] proposed a graph-based regularization technique, formulating a global optimization challenge to simultaneously upgrade the resolution across all LF views. This process involves rough disparity estimation for warping calculations and geometric structuring among views for optimized super-resolution outcomes. However, despite the advent of numerous disparity estimation techniques [45], the reconstructed view images remain prone to errors. These inaccuracies often lead to noticeable artifacts, especially near occlusion boundaries, highlighting the challenges still faced in achieving accurate LFSR.

2.2. CNN-based Methods

Recently, advancements have shown that deep Convolutional Neural Networks (CNNs) outperform traditional techniques in enhancing the spatial resolution of Light Fields (LF). In the groundbreaking study, LFCNN [21], Sub-Aperture Images (SAIs) were initially improved in resolution using SRCNN [46], followed by a fine-tuning process in pairs to boost both spatial and angular sharpness. Building on this, Yuan refined the LFCNN approach by employing EDSR [47] for each SAI's super-resolution, alongside crafting an EPI-enhancement network to refine preliminary outcomes. Jin [22] introduced a novel strategy for spatial super-resolution by advocating an all-to-one approach, incorporating structural consistency regulation to safeguard parallax integrity. Meanwhile, Zhang et al. [24] developed a multi-stream residual framework, utilizing stacks of SAIs from varying angular perspectives as input. Extending their work, Zhang et al. [23] and associates further amplified super-resolution efficacy through the execution of 3D convolutions on these SAI stacks across different angular orientations. Yeung et al. [29] proposed LFSSR to alternately reshape LF images between the SAI pattern and MacPI pattern for convolution. Most recently, Wang et al. [48] have leveraged deformable convolution techniques on LF imagery, specifically to navigate and rectify the disparity challenges inherent in LF spatial super-resolution. More recently, Wang et al. [27] introduced the DistgSSR network, bringing to the fore a novel and efficient mechanism for disentangling complex image features. This approach

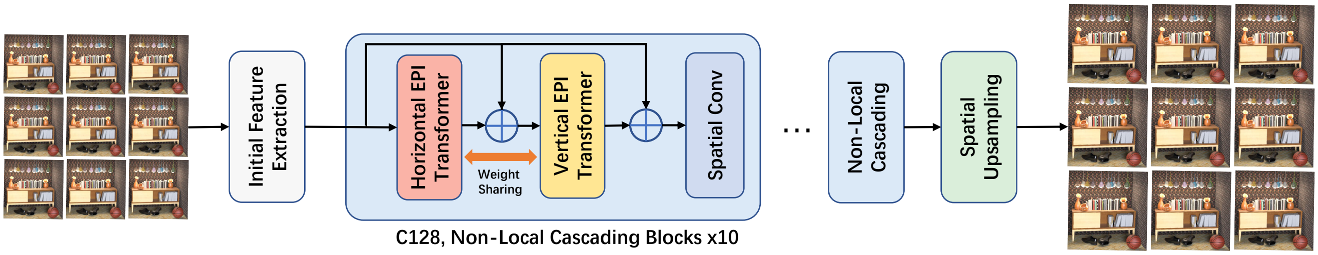

Figure 2. An overview of our BigEPIT network. A 3×3 LF is used as an example for illustration.

employs distinct Spatial, Angular, and Epipolar Plane Feature Extractors, each tailored to isolate and process unique aspects of LF data. This methodology underscores a significant advancement in precision and quality of LF image super-resolution, enabling more detailed and accurate reconstruction of light field imagery.

2.3. Transformer-based Methods

Lately, transformer-based models, celebrated for their capacity to handle long-range dependencies, have found extensive application across a variety of visual tasks. Models like the Vision Transformer (ViT) [49] for image classification, DETR [50] for object detection, and SETR [51] for semantic segmentation have recorded remarkable achievements in foundational tasks of computer vision. Their success underscores the transformative impact of transformer architecture in enhancing the accuracy and efficiency of computer vision applications. DPT [52] and LFT [26] stand out as two notable examples of employing transformer technology for LF-SSR. DPT utilizes transformers to treat SAIs along each vertical or horizontal axis as sequences, delving into the long-range relationships within them. On the other hand, LFT adopts a dual approach, alternating between angular transformers, which focus on modeling each macro-pixel, and spatial transformers, dedicated to each SAI. This strategy effectively merges angular and spatial insights, showcasing the versatility and depth of transformers in enhancing LF-SSR. More recently, EPIT [1] has delved further into addressing the challenges posed by large disparity variations inherent in LF. It specifically models the long-range dependencies present within EPIs, showcasing an advanced approach to understanding and managing the complexities associated with LF imaging. This innovation marks a significant step forward in enhancing the precision and effectiveness of LF image analysis.

3. Method

In this section, we first introduce the network architecture and then present some improvements, including compound model scaling, augmented data resampling, and a high-precision test scheme.

3.1. Network Architecture

An overview of our BigEPIT is shown in Fig. 2. Our network takes an LR $\mathcal{L}_{LR} \in \mathbb{R}^{U\times V\times H\times W}$ as its input and produces an HR LF $\mathcal{L}_{HR} \in \mathbb{R}^{U\times V\times \alpha H\times \alpha W}$ where α presents the upscaling factor. Our network consists of three stages including initial feature extraction, non-local cascading block, and spatial upsampling. Please refer to EPIT [1] for more details.

Initial Feature Extraction: We follow the EPIT [1] to use three 3×3 convolutions with LeakyReLU to map each SAI to a high-dimensional feature. The initially extracted feature can be represented as $F \in \mathbb{R}^{U\times V\times H\times W\times C}$, where C denotes the channel dimension.

Non-Local Cascading Block: Through stacking several of the Non-Local Cascading blocks, the model can achieve a global perception of all angular views and follow SwinIR [53] to adopt spatial convolutions to enhance the local feature representation. The non-local cascading block consists of a horizontal EPI transformer, a vertical EPI transformer, and spatial convolutions sequentially. Note that the weights of the two basic transformer units in each block are shared, which can help teach the spatial-angular correlation better.

Spatial Upsampling: Following EPIT [1], 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 work. We also employ L1 loss function to train our network due to its robustness to outliers. We convert input images into the YCbCr color space, and only super-resolve the Y channel of images, leaving Cb and Cr channel images being bicubicly upscaled.

3.2. Compound Model Scaling

Drawing inspiration from EfficientNet [40], the fundamental idea is that deeper networks possess the ability to capture complex and richer features, leading to improvement in the SR task. However, training deep networks can be challenging due to the problem of vanishing gradients. On the other hand, wider networks can capture finer details and are comparatively easier to train. Nevertheless, extremely wide but shallow networks often struggle to capture higher-level features. In summary, we adopt a compound scaling strategy,

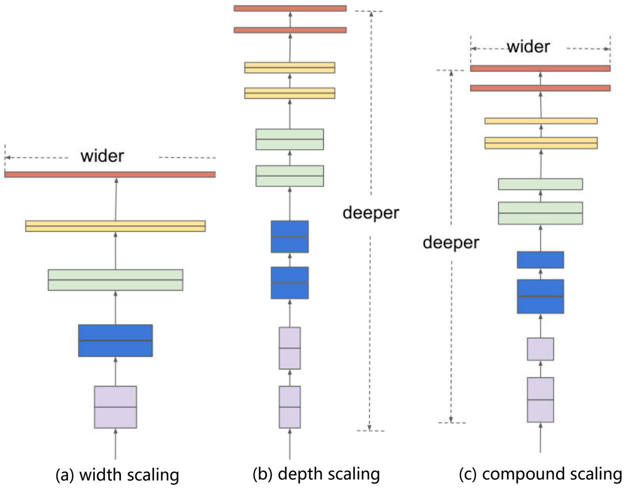

Figure 3. Different variations of model scaling. (a) Width scaling. (b) Depth scaling. (c) Compound scaling that uniformly scales width and depth with a fixed ratio. Figure references from [40].

which involves simultaneously scaling the width and depth of the network with a fixed ratio, e.g., 2. Specifically, we increase the width from **64** → **128** and the number of the Non-Local Cascading block from **5** → **10**.

3.3. Augmented Data Resampling

We find that in the existing LF datasets, there is an imbalance in disparity and large disparity data are scarce, which causes the model to be sensitive to large disparity. Assume that LFs have an angular resolution of 9×9. Inspired by [41], we also use the augmented data sampling strategy to extract 5×5 SAIs for training and testing, as shown in Fig. 4, including central sampling, even sampling, and uneven sampling. This strategy explicitly increases the number of images of large disparity, which can improve the robustness of the model to disparity variations, while increasing the training time. Therefore, a trade-off needs to be considered in terms of time and precision.

3.4. High-Precision Test Scheme

While BigEPIT effectively learns disparity features, the traditional post-processing test scheme that employs center padding disrupts the disparity structural correlation within the SAI subspace, as illustrated in Fig. 5 (a). Specifically, the SAIs captured by LF cameras adhere to strict optical disparity constraints. Each subspace within the LF image exhibits significant spatial-angular correlation, with the disparity values gradually decreasing from the outermost layer towards the center. However, the introduction of artifact padding values with an inaccurate disparity structure through center padding undermines the disparity relationship within the subspace, leading to poorer predictions

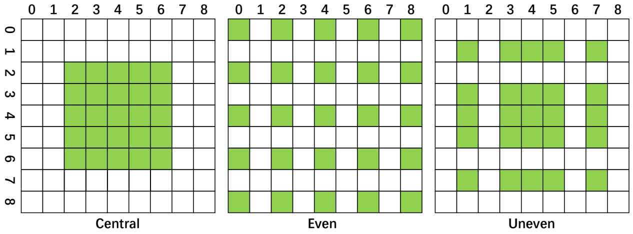

Figure 4. Illusatrtion of the augmented data sampling strategy, including central sampling, even sampling, and uneven sampling.

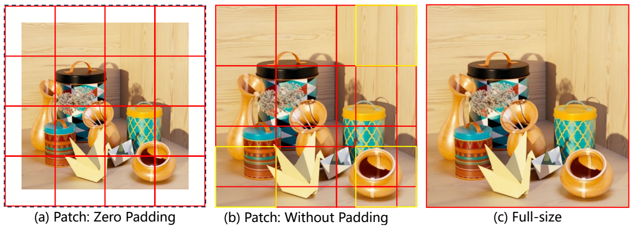

Figure 5. Illustrations of different LF cropped image methods in the testing phase. Here, the center SAI is used for illustration.

by the networks. To address this issue, DistgEPIT [41] proposed a Position-Sensitive Windowing (PSW) operation aimed at preserving the disparity structural consistency within the SAI (Sparse Angular Image) subspace during windowing, without requiring additional computations. Without the need for padding operations, this operation utilizes a sliding window approach to crop the block in an overlapping manner. For border values, it backtracks to fill in the entire block, as depicted in Fig. 5 (b).

However, existing patch-based testing schemes often employ smaller patches, e.g., 32×32. Through experiments, we found that our model's performance follows the principle of "the larger the image patch size, the better." In the extreme case of directly inputting the entire image, it achieves the best performance. We analyze this phenomenon as being attributed to our model's ability to obtain a sufficiently large receptive field, allowing for better utilization of spatial and angular correlations. Therefore, we suggest implementing a high-precision test scheme, i.e., full-size or larger patch size is employed without zero padding whenever the available GPU memory permits to achieve superior results.

4. Experiments

In this section, we first introduce the dataset, implementation details, and evaluation metrics. Then, we quantitatively and qualitatively compare our approach with state-of-the-art methods. Next, we compare the performance of different LF image SR methods in real LF scenes. Finally, we verify the effectiveness and robustness of our method through a series of ablation experiments.

Table 1. PSNR/SSIM metrics comparison among the other prestigious approaches for 4× upscaling factors with 5×5 angulars resolution. The best averaged results are achieved by our BigEPIT method highlighted in bold fonts. For a fair comparison, we also use the Position-Sensitive Windowing (PSW) test scheme setting with DistgEPIT [41]. Aug means trained on an augmented data resampling strategy, and TTA means the test-time-augmentation technique by using seven different affine transformations.

Methods	EPFL	HCInew	HCIold	INRIA	STFgantry	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[54]	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[47]	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[55]	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[24]	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[29]	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[22]	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[28]	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[23]	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[56]	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[52]	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[52]	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[27]	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[1]	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 [41]	30.09 / 0.9224	31.61 / 0.9252	37.96 / 0.9742	32.35 / 0.9535	32.45 / 0.9589	32.90 /0.9468
BigEPIT	**30.26 / 0.9236**	**31.80 / 0.9264**	**38.05 / 0.9754**	**32.40 / 0.9547**	**32.70 / 0.9601**	**33.04 / 0.9480**
DistgEPIT+Aug [41]	30.17 / 0.9232	31.71 / 0.9263	38.03 / 0.9744	32.39 / 0.9535	32.74 / 0.9604	33.01 / 0.9476
BigEPIT+Aug	**30.36 / 0.9256**	**31.85 / 0.9270**	**38.08 / 0.9750**	**32.51 / 0.9557**	**33.00 / 0.9619**	**33.16 / 0.9491**
DistgEPIT+Aug+TTA [41]	30.41 / 0.9260	31.91 / 0.9283	38.24 / 0.9753	32.60 / 0.9551	33.06 / 0.9626	33.25 / 0.9495
BigEPIT+Aug+TTA	**30.61 / 0.9282**	**32.05 / 0.9287**	**38.31 / 0.9761**	**32.76 / 0.9573**	**33.27 / 0.9636**	**33.40 / 0.9507**

4.1. Datasets and Implementation Details

In our LF image SR experiments, we utilize five widely used LF image datasets: EPFL [58], HCINew [59], HCI-old [42], INRIA [60], and STFgantry [61], following the approach of previous methods [26, 28, 48, 52, 56]. All LFs in these datasets have an angular resolution of 9 × 9. For training, we employ the augmented data sampling strategy, as depicted in Fig. 4, to extract 5 × 5 SAIs. For testing, we only extract the central 5 × 5 SAIs. During the training stage, each SAI is cropped into patches of size 128 × 128 with a stride of 32. LF patches of size 32 × 32 are generated using bicubic downsampling. To augment the training data, we apply random horizontal flipping, vertical flipping, and 90-degree rotation. The Adam optimizer [62] is employed with a batch size of 2 per GPU and a learning rate of 2×10^{-4}, which is halved every 15 epochs. Our BigEPIT model with augmented data sampling is implemented in the PyTorch framework and trained for 30 epochs using four Nvidia A100 GPUs. Similarly, our BigEPIT model with central sampling is trained for 50 epochs using two Nvidia A100 GPUs. For evaluation, we employ the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [63] as the evaluation metrics. Following the procedure of previous methods [26, 28, 48, 52, 56] when calculating the average metric for a dataset, we first average the metric scores for all SAIs within each scene, and then average the metric scores across all scenes in the dataset.

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

We compare our method with 15 state-of-the-art SR methods, including 3 single image SR methods [47, 54, 55] and 12 LF image SR methods [1, 22–24, 26–29, 41, 48, 52, 56].

Quantitative Results. Table 1 shows a quantitative comparison among LF image SR methods. Our BigEPIT of different versions all achieves state-of-the-art quantitative metrics, i.e., the final version is PSNR of **33.40 dB** and SSIM of **0.9507**, exceeding 0.15 dB compared to DistgEPIT [41]. Note that for a fair comparison, we use the PSW test scheme setting with DistgEPI [41]. This demonstrates the robustness of our BigEPIT, in particular the large disparity dataset, the STFgantry [61], captured with a Lego Gantry.

Qualitative Results. Figure 6 shows the qualitative results achieved by different methods for 4× SR. As can be seen from the zoomed-in areas, the SISR method (i.e., VDSR) cannot reliably recover the missing details, such as the handrail area in the scene *Sculpture*. While other LFSR methods have achieved promising results, they often struggle with reconstructing numerical patterns in the scene *ISO Chart*, and the corresponding texture in EPI images is not sufficiently clear. In contrast, our approach not only achieves superior visual quality but also exhibits sharper and clearer lines with fewer artifacts in EPI images. The

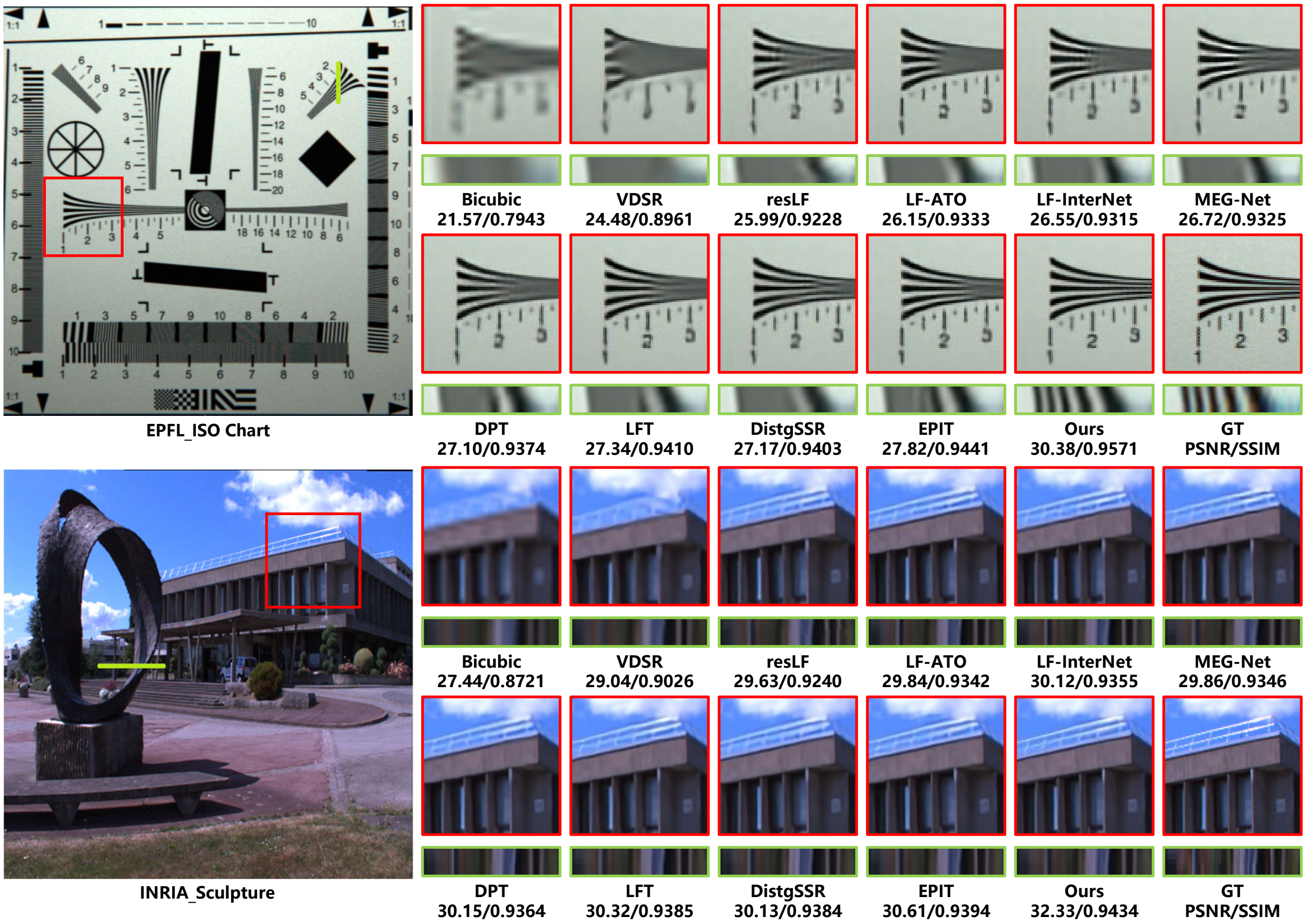

Figure 6. Qualitative comparison of different SR methods for 4×SR. The super-resolved center view images, vertical or horizontal EPIs are shown. Best viewed zoom-in electronically.

Table 2. Comparisons of PSNR metric among different model scaling schemes in EPIT [1] for 4× SR, with the best results highlighted in bold. FLOPs are computed with an input LF of size 5×5×32×32. Note that all methods are trained and tested with the same settings.

Scaling	Channels	Blocks	#Params.	Flops	EPFL	HCInew	HCIold	INRIA	STFgantry	Average
base	64	5	1.14M	55.30G	29.81	31.45	37.73	32.05	32.07	32.62
width	128	5	4.55M	216.19G	29.97	31.57	37.83	32.21	32.25	32.77
width	180	5	9.02M	426.95G	30.02	31.63	37.86	32.20	32.30	32.80
depth	64	10	2.11M	103.85G	29.95	31.58	37.80	32.19	32.23	32.75
depth	64	20	4.04M	202.94G	29.99	31.58	37.85	32.23	32.27	32.79
depth	64	45	8.86M	450.67G	30.04	31.65	37.88	32.22	32.32	32.82
compound	128	10	8.32M	413.85G	**30.25**	**31.80**	**38.05**	**32.40**	**32.70**	**33.04**

slope of these lines is related to the depth values, which demonstrates that our method can preserve the LF disparity structure well.

Angular Consistency. Furthermore, we assess the angular consistency by employing a depth estimation algorithm called SPO [57]. As depicted in Fig. 7, the depth estimation results obtained using the SISR method exhibit significant noise and higher MSE ×100 error, indicating a lack of consideration for the angular information in LF images. In contrast, when utilizing LF images generated by our method, we observe a substantial improvement in depth estimation accuracy. The resulting depth map exhibits visually clear quality and demonstrates the lowest MSE ×100 error compared to alternative methods. This further highlights the su-

Table 3. Comparisons of PSNR metric among different training data resampling for 4× SR, with the best results highlighted in bold.

Resampling	EPFL	HCInew	HCIold	INRIA	STFgantry	Average
Central	30.25	31.80	38.05	32.40	32.70	33.04
Central+Even	30.34	31.84	38.07	32.49	32.91	33.13
Central+Even+Uneven	**30.36**	**31.85**	**38.08**	**32.51**	**33.00**	**33.16**

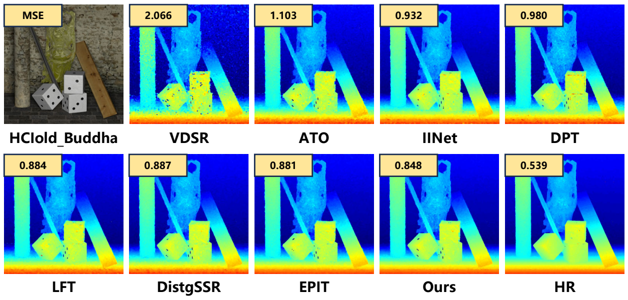

Figure 7. Depth estimation results achieved by SPO [57] using 4× SR LF images produced by different SR methods. We report the mean square error multiplied by 100 (MSE ×100). Lower is better.

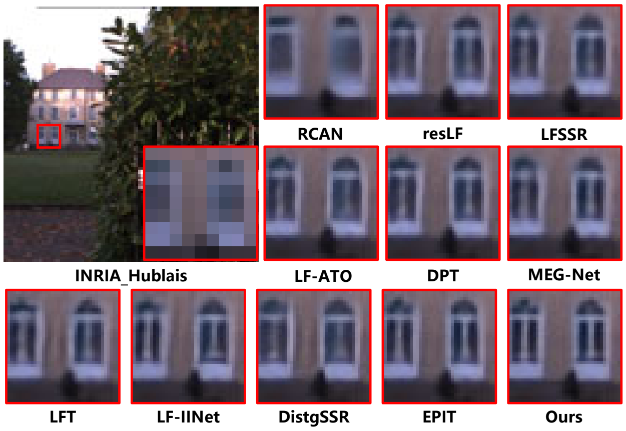

Figure 8. Visual comparison of different SR methods on real-world LF scene for 4× SR.

perior angular consistency achieved by our method.

Performance on Real Scenes. We conduct a comparison of different methods on the real scene *Hublais* from the INRIA dataset [60]. As illustrated in Fig. 8, certain SR methods, such as RCAN [55] and resLF [24], yield blurry results lacking realism. Similarly, the outcomes obtained from InterNet [48] and DistgSSR [27] methods exhibit noticeable unrealistic artifacts. In contrast, our method generates results with clearer edges and enhanced visual perception compared to the other methods. This observation confirms that our approach successfully achieves the intended goals with superior robustness.

4.3. Ablation Study

In this subsection, we further analyze the effectiveness of the strategies proposed in our work, including compound model scaling, augmented data resampling, and a high-precision test scheme.

Model Scaling. We perform scaling on the EPIT model based on different strategies, i.e., width scaling, depth scaling, and compound scaling. In Table 2, we first validate the effectiveness of width scaling. We observe that increasing the width from 64 to 128 results in a PSNR improvement of 0.15 dB. However, when the width is further increased from 128 to 180, the PSNR only increases by 0.03 dB. This indicates that solely increasing the width leads to a quick saturation of the model's performance. Similar observations can be made for depth scaling. Furthermore, we adjust the corresponding number of channels and blocks to ensure that the models following the three scaling strategies have roughly the same number of parameters and Flops. From Table 2, it can be observed that our compound scaling achieves the best results. This demonstrates the importance of considering both width and depth when scaling the model.

Study of Training Data. We continue our investigation into the impact of training data resampling strategies on the model. As shown in Table 3, the utilization of Even sampling significantly enhances the model's performance on the large-disparity dataset (STFgantry [61]), resulting in a notable increase of 0.21 dB. Corresponding improvements can also be observed on other datasets. Moreover, when combined with uneven sampling, the performance can be further improved. Finally, the augmented data resampling can achieve an average PSNR improvement of 0.16 dB. It is important to note that employing the augmented resampling strategy increases training time and requires a trade-off between training time and model accuracy.

Study of Test Schemes. Lastly, we investigate the impact of testing schemes on the model performance, taking pre-trained EPIT as an example. From the table, it can be observed that PSW without zero padding improves the PSNR by 0.23 dB, and the performance improvement becomes more significant as the tested patch size increases. The best results are obtained when testing with the full-size input, which elevates the PSNR of the EPIT model from 32.42 dB to 32.71 dB. This demonstrates the importance of selecting an appropriate testing scheme. Therefore,

Table 4. Quantitative PSNR comparison among different test strategies and settings of pre-trained EPIT [1] model. † means using PSW strategy due to limited GPU memory.

Scheme	Patchsize	Stride	Padding	EPFL	HCInew	HCIold	INRIA	STFgantry	Average
Origin	32	16	Zero	29.34	31.51	37.68	31.37	32.18	32.42
PSW	32	16	-	29.85	31.46	37.67	32.11	32.16	32.65
PSW	48	24	-	29.87	31.50	37.69	32.12	32.19	32.67
PSW	64	32	-	29.88	31.51	37.70	32.15	32.22	32.69
PSW	96	48	-	29.89	31.52	37.72	32.15	32.24	32.70
Full-size	-	-	-	**29.90**	**31.53**	**37.72**†	**32.16**	**32.24**†	**32.71**

Table 5. Our team achieved first place in the leaderboard (last three rows) on the NTIRE-2024 test dataset. TTA means the test-time-augmentation technique by using seven different affine transformations.

Methods	Params.	Lytro	Synthetic	Average
Bicubic	-	25.109 / 0.8404	26.461 / 0.8352	25.785 / 0.8378
VDSR[54]	0.665 M	27.052 /0.8888	27.936 / 0.8703	27.494 / 0.8795
EDSR[47]	38.89 M	27.540 / 0.8981	28.206 / 0.8757	27.873 / 0.8869
RCAN[55]	15.36 M	27.606 / 0.9001	28.308 / 0.8773	27.957 / 0.8887
resLF[24]	8.646 M	28.657 / 0.9260	29.245 / 0.8968	28.951 / 0.9114
LFSSR[29]	1.774 M	29.029 / 0.9337	29.399 / 0.9008	29.214 /0.9173
LF-ATO[22]	1.364 M	29.087 / 0.9354	29.401 / 0.9012	29.244 / 0.9183
LF-InterNet[28]	5.483 M	29.233 / 0.9369	29.446 / 0.9028	29.340 / 0.9198
MEG-Net[56]	1.775 M	29.203 / 0.9369	29.539 / 0.9036	29.371 / 0.9203
LF-IINet[52]	4.886 M	29.487 / 0.9403	29.786 / 0.9071	29.636 / 0.9237
DPT[52]	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[27]	3.582 M	29.389 / 0.9403	29.884 /0.9084	29.637 / 0.9244
EPIT[1]	1.470 M	29.718 /0.9420	30.030 /0.9097	29.874 / 0.9259
HLFSR-SSR[35]	13.87 M	29.714 /0.9429	29.945 / 0.9097	29.830 / 0.9263
DistgEPIT[41]-TTA	19.02M	30.746 / 0.9468	30.460 / 0.9146	30.603 / 0.9307
BigEPIT-TTA	8.32M	30.951 / 0.9492	30.578 / 0.9164	30.765 / 0.9328
BNU&TMU-AI-TRY	/	**31.003 / 0.9496**	**30.602 / 0.9167**	**30.803 / 0.9332**
BITSMBU	/	30.930 / 0.9486	30.525 / 0.9159	30.727 / 0.9322
OpenMeow	/	30.961 / 0.9491	30.457 / 0.9154	30.709 / 0.9323

when GPU resources allow, testing with the full-size input or larger patch sizes without zero padding can be employed to achieve higher accuracy.

4.4. NTIRE 2024 LFSR Challenge Results

In the NTIRE 2024 LFSR challenge, a new dataset is developed, consisting of 16 synthetic LFs and 16 real-world LFs captured by the Lytro camera for the test subset. During the challenge, participants are strictly prohibited from using any external models or data, including pre-trained backbones and optical flow networks. For reporting the final results, we employed the average ensemble method to combine the outputs generated by BigEPIT, DistgEPIT [41], and RR-HLFSR [35] with TTA, since the three models have different structures. Table 5 shows that our team achieved 1st place with a PSNR of **30.803** dB on the LFSR test dataset.

5. Conclusion

In this paper, we have proposed a series of strategies that successfully scale the EPIT model to BigEPIT, which can better solve the disparity problem in LF image SR, including compound model scaling, augmented data resampling, and a high-precision test scheme. Experimental results show that our model can yield visually pleasant and angular consistent SR results on synthetic and real-world LF images, and achieved 1st place in the NTIRE 2024 Light Field Image Super-Resolution Challenge.

References

[1] 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. *arXiv preprint arXiv:2302.08058*, 2023. 1, 3, 5, 6, 8

[2] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4748–4757, 2018. 1

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

[4] Wentao Chao, Fuqing Duan, Xuechun Wang, Yingqian Wang, and Guanghui Wang. Occcasnet: occlusion-aware cascade cost volume for light field depth estimation. *arXiv preprint arXiv:2305.17710*, 2023.

[5] Jiaxin Chen, Shuo Zhang, and Youfang Lin. Attention-based multi-level fusion network for light field depth estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1009–1017, 2021.

[6] Numair Khan, Min H Kim, and James Tompkin. Differentiable diffusion for dense depth estimation from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8912–8921, 2021.

[7] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2022.

[8] Titus Leistner, Radek Mackowiak, Lynton Ardizzone, Ullrich Köthe, and Carsten Rother. Towards multimodal depth estimation from light fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12953–12961, 2022.

[9] Hao Sheng, Yebin Liu, Jingyi Yu, Gaochang Wu, Wei Xiong, Ruixuan Cong, Rongshan Chen, Longzhao Guo, Yanlin Xie, Shuo Zhang, et al. Lfnat 2023 challenge on light field depth estimation: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3472–3484, 2023.

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

[11] Gaochang Wu, Yebin Liu, Lu Fang, Qionghai Dai, and Tianyou Chai. Light field reconstruction using convolutional network on epi and extended applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1681–1694, 2018. 1

[12] 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. 1

[13] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus H Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Transactions on Graphics (TOG)*, 32(4):73–1, 2013. 1

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

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

[16] 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. IEEE, 2012.

[17] 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.

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

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

[20] Mattia Rossi and Pascal Frossard. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Transactions on Image Processing*, 27(9):4207–4218, 2018. 1, 2

[21] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24(6):848–852, 2017. 1, 2

[22] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2260–2269, 2020. 2, 5, 8

[23] Shuo Zhang, Song Chang, and Youfang Lin. End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Transactions on Image Processing*, 30:5956–5968, 2021. 2, 5

[24] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11046–11055, 2019. 2, 5, 7, 8

[25] 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.

[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. 1, 3, 5, 8

[27] 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. 2, 5, 7, 8

[28] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *ECCV 2020*, pages 290–308. Springer, 2020. 5, 8

[29] Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018. 2, 5, 8

[30] Nan Meng, Hayden K-H So, Xing Sun, and Edmund Y Lam. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):873–886, 2019.

[31] 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*, pages 11757–11764, 2020.

[32] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10010–10019, 2021.

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

[34] Yangling Chen, Shuo Zhang, Song Chang, and Youfang Lin. Light field reconstruction using efficient pseudo 4d epipolar-aware structure. *IEEE Transactions on Computational Imaging*, 8:397–410, 2022.

[35] Vinh Van Duong, Thuc Nguyen Huu, Jonghoon Yim, and Byeungwoo Jeon. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Transactions on Computational Imaging*, 9:350–366, 2023. 1, 8

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

[37] Wentao Chao, Fuqing Duan, Xuechun Wang, Yingqian Wang, and Guanghui Wang. Lfsrdiff: Light field image super-resolution via diffusion models. *arXiv preprint arXiv:2311.16517*, 2023.

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

[39] Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte, Yulan Guo, et al. Ntire 2024 challenge on light field image super-resolution: Methods and results. In *CVPRW*, 2024. 1

[40] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning (ICML)*, pages 6105–6114. PMLR, 2019. 2, 3, 4

[41] Kai Jin, Angulia Yang, Zeqiang Wei, Sha Guo, Mingzhi Gao, and Xiuzhuang Zhou. Distgepit: Enhanced disparity learning for light field image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1373–1383, 2023. 2, 4, 5, 8

[42] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, volume 13, pages 225–226. Citeseer, 2013. 2, 5

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

[44] Shuo Zhang, Hao Sheng, Da Yang, Jun Zhang, and Zhang Xiong. Micro-lens-based matching for scene recovery in lenslet cameras. *IEEE Transactions on Image Processing*, 27(3):1060–1075, 2017. 2

[45] Ole Johannsen, Katrin Honauer, Bastian Goldluecke, Anna Alperovich, Federica Battisti, Yunsu Bok, Michele Brizzi, Marco Carli, Gyeongmin Choe, Maximilian Diebold, et al. A taxonomy and evaluation of dense light field depth estimation algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 82–99, 2017. 2

[46] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 2

[47] 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. 2, 5, 8

[48] Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo. Light field image super-resolution using deformable convolution. *IEEE Transactions on Image Processing*, 30:1057–1071, 2020. 2, 5, 7

[49] 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. *arXiv preprint arXiv:2010.11929*, 2020. 3

[50] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 3

[51] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021. 3

[52] Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Detail preserving transformer for light field image super-resolution. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 3, 5, 8

[53] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

[54] 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

[55] 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, 7, 8

[56] 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

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

[58] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 5

[59] 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. Springer, 2016. 5

[60] Mikael Le Pendu, Xiaoran Jiang, and Christine Guillemot. Light field inpainting propagation via low rank matrix completion. *IEEE Transactions on Image Processing*, 27(4):1981–1993, 2018. 5, 7

[61] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7):3, 2008. 5, 7

[62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[63] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5