NTIRE 2024 Challenge on Light Field Image Super-Resolution: Methods and Results

Yingqian Wang*, Zhengyu Liang*, Qianyu Chen*, Longguang Wang*, Jungang Yang*†, Radu Timofte*, Yulan Guo*, Wentao Chao, Yiming Kan, Xuechun Wang, Fuqing Duan, Guanghui Wang, Wang Xia, Ziqi Wang, Yue Yan, Peiqi Xia, Shunzhou Wang, Yao Lu, Angulia Yang, Kai Jin, Zeqiang Wei, Sha Guo, Mingzhi Gao, Xiuzhuang Zhou, Zhongxin Yu, Shaofei Luo, Cheng Zhong, Shaorui Chen, Long Peng, Yuhong He, Gaosheng Liu, Huanjing Yue, Jingyu Yang, Zhengjian Yao, Jiakui Hu, Lujia Jin, Zhi-Song Liu, Chenhang He, Jun Xiao, Xiuyuan Wang, Zonglin Tian, Yifan Mao, Deyang Liu, Shizheng Li, Ping An

Abstract

In this report, we summarize the 2nd NTIRE challenge on light field (LF) image super-resolution (SR) with a focus on new methods and results. This challenge aims at super-resolving LF images under the standard bicubic downsampling degradation with a magnification factor of $\times 4$. Compared with single image SR, the major challenge of LF image SR lies in how to exploit complementary angular information from plenty of views with varying disparities. This year of challenge has two tracks, including one track on fidelity (i.e., restoration accuracy in terms of PSNR) only, and the other track on fidelity with an extra constraint on model size and computational cost. In total, 125 participants were successfully registered for this challenge, and 9 teams have successfully submitted results with PSNR scores higher than the baseline methods. We report the solutions proposed by the participants, and summarize their common trends and useful tricks. We hope this challenge can stimulate future research and inspire new ideas in LF image SR.

1. Introduction

Light field (LF) cameras are capable of capturing the intensity and directions of light rays, allowing for the recording of 3D geometry in a practical and effective way. Through encoding 3D scene cues into 4D LF images (2D for spatial dimension and 2D for angular dimension), LF cameras can facilitate numerous appealing applications, including post-capture refocusing [1, 2], depth sensing [3–5], virtual reality [6, 7], and view rendering [8–11].

In many applications, there is a significant need for high-resolution (HR) LF images in order to achieve enhanced perceptual quality and provide advantages for subsequent applications. Nonetheless, obtaining HR LF images typically comes at a considerable cost due to the inherent spatial-angular trade-off problem in LF imaging. Therefore, it is crucial to reconstruct HR LF images from their low-resolution (LR) counterparts, a process known as LF image super-resolution (SR).

Recently, significant advancements have been made in image SR by utilizing deep learning methods. Nevertheless, the majority of these approaches concentrate on enhancing the resolution of single images [12–19], stereo images [20–23], or videos [24, 25], and are not readily applicable to the task of LF image SR. When dealing with LF images, effectively integrating both spatial and angular information is crucial yet challenging due to the unique structure of LF data.

To develop and benchmark LF image SR methods, the 1st LF image SR challenge was hosted on the NTIRE 2023 workshop [26]. This challenge employed the widely used and publicly available LF datasets [27–31] as training set, and proposed a new LF dataset for both validation (model development) and test (final ranking). The popular bicubic downsampling degradation is used to generate LR LF images, and the objective of this challenge is to make the

*Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte and Yulan Guo are the NTIRE 2024 challenge organizers, while the other authors participated in this challenge.
†Corresponding author: Jungang Yang
Section 6 provides the authors and affiliations of each team.
NTIRE 2024 webpage: https://cvlai.net/ntire/2024/
Challenge webpage (Track 1): https://codalab.lisn.upsaclay.fr/competitions/17265
Challenge webpage (Track 2): https://codalab.lisn.upsaclay.fr/competitions/17266
Github: https://github.com/The-Learning-And-Vision-Atelier-LAVA/LF-Image-SR/tree/NTIRE2024
BasicLFSR toolbox: https://github.com/ZhengyuLiang24/BasicLFSR

super-resolved LF images as faithful as the groundtruth HR ones. However, an important issue in image SR, *i.e.*, the computational efficiency, was not considered in the 1st LF image SR challenge.

Succeeding the previous year, we hold the 2nd LF image SR challenge on the NTIRE workshop in 2024. This challenge has two competition tracks. Track 1 is inherited from the NTIRE 2023 challenge, focusing on the restoration fidelity (*i.e.*, PSNR) only. Track 2 not only focuses on the restoration fidelity, but also has a strict constraint on model size (*i.e.*, the number of parameters) and computational cost (*i.e.*, FLOPs). We introduce Track 2 in order to inspire the community to explore the specific challenges in model deployment, and stimulate the research for practical LF image SR.

This challenge is one of the NTIRE 2024 workshop associated challenges on: dense and non-homogeneous dehazing [32], night photography rendering [33], blind compressed image enhancement [34], shadow removal [35], efficient super resolution [36], image super resolution ($\times$4) [37], light field image super-resolution [38], stereo image super-resolution [39], HR depth from images of specular and transparent surfaces [40], bracketing image restoration and enhancement [41], portrait quality assessment [42], quality assessment for AI-generated content [43], restore any image model (RAIM) in the wild [44], RAW image super-resolution [45], short-form UGC video quality assessment [46], low light enhancement [47], and RAW burst alignment and ISP challenge.

2. Related Work

In this section, we will provide a brief overview of significant works in the field of LF image super-resolution (SR). We categorize existing LF image SR methods into three main groups: traditional (i.e., non-learning) methods, CNN-based methods, and Transformer-based methods.

2.1. Traditional Methods

Light field image SR has been a long-standing research challenge and has been investigated for decades. Bishop et al. [48] proposed a Bayesian deconvolution approach to super-resolve LF images based on the estimated disparities. Wanner et al. [49] initially estimated disparity maps using structure tensor, and then developed a variational framework for LF image SR. Farrugia et al. [50] constructed a patch-volume dictionary of HR and LR LF image pairs, and introduced a multivariate ridge regression method to learn the linear mapping from LR patch volumes to their HR counterparts. Alain et al. [51] addressed the ill-posed LF image SR problem as an optimization problem based on the sparsity prior. Rossi et al. [52] integrated inter-view information using graph regularization and formulated LF image SR as a quadratic problem, which can be efficiently solved with standard convex optimization techniques.

2.2. CNN-based Methods

In the last decade, convolutional neural networks (CNNs) have been extensively researched and have shown remarkable performance in LF image SR. Yoon et al. [53] introduced the first CNN-based LF image SR method, known as LFCNN. In their method, input LF images were grouped into pairs or quads and passed through a three-layer CNN to integrate complementary information from neighboring views. This pioneering work demonstrated the potential of CNNs in LF image SR. Since then, numerous deeper CNN architectures with various mechanisms for incorporating angular information have been developed to achieve improved SR performance in LF image SR tasks.

Wang et al. [54] proposed a bidirectional recurrent CNN to incorporate angular information from the sub-aperture images (SAIs) along the horizontal or vertical angular direction. Zhang et al. [55] stacked SAIs along four different angular directions and developed a four-branch residual network to implicitly learn the epipolar geometry from stacked SAIs for LF image SR. In their subsequent work, Zhang et al. [56] improved SR performance by performing 3D convolutions on SAI stacks of different angular directions. Cheng et al. [57] developed a framework to exploit both internal and external similarities for LF image SR. Meng et al. [58] applied 4D convolutions to simultaneously incorporate spatial and angular information from 4D LF data and developed the high-dimensional dense residual network (HDDRNet) for LF image SR. Jin et al. [59] proposed an all-to-one method for LF image SR and performed structural consistency regularization to preserve the parallax structure. Wang et al. [60] applied deformable convolution to LF spatial SR and designed a collect-and-distribute scheme to incorporate complementary information from different views. Mo et al. [61] proposed a dense dual-attention network (DDAN) for LF image SR, which included a view attention module and a channel attention module to adaptively capture discriminative information from different views and channels.

Additionally, some methods decoupled high-dimensional LF data into different subspaces for LF image SR. Yeung et al. [62] alternately reshaped LF images between the SAI pattern and macro-pixel pattern, and designed spatial-angular separable convolutions for LF image SR. Wang et al. [63] proposed spatial and angular feature extractors to extract corresponding information from macro-pixel images and developed an LF-InterNet to repetitively interact spatial and angular information for LF image SR. In their subsequent work [64], Wang et al. further generalized the interaction mechanism into

https://cvlai.net/ntire/2024/

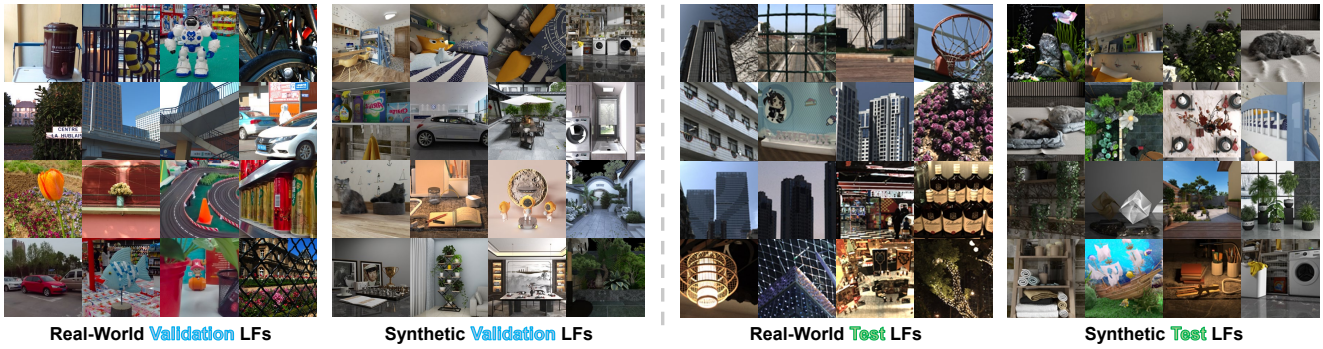

Figure 1. An illustration of the center-view images in the NTIRE-LFSR dataset [26]. Both validation and test sets contain 16 real-world and 16 synthetic LFs, respectively.

an LF disentangling mechanism, and developed three CNNs (DistgSSR, DistgASR, and DistgDisp) for spatial SR, angular SR, and disparity estimation, respectively. Following LF-InterNet [63], Liu et al. [65] proposed an intra-inter view interaction network (LF-IINet) with two parallel branches to extract global inter-view information and model correlations among all intra-view features. These two branches are mutually interacted to fuse angular and spatial information for LF image SR.

Besides the aforementioned works that design advanced network structures to pursuit superior SR accuracy, several works also studied some special yet important issue in LF image SR. Cheng et al. [66] addressed the domain gap issue by proposing a "zero-shot" learning framework, in which the network learns to achieve spatial SR without using external training data except the given input LR LF. Wang et al. [67] addressed the degradation formulation issue in LF image SR, and proposed a method to handle LF image SR with multiple degradation. Xiao et al. [68] proposed a data augmentation approach tailored for LF image SR, which can be applied to existing LF image SR networks to further improve their SR performance.

2.3. Transformer-based Methods

Transformer networks, which were originally developed for natural language processing [69], have recently gained much attention in computer vision community. Recently, Transformers have been successfully applied to many low-level vision tasks such as image restoration [18, 70, 71] and video SR [72–74], and achieved superior performance than CNN-based methods.

In the past two years, researchers have explored Transformers for LF image SR. Wang et al. [75] proposed a detail-preserving Transformer (DPT) for LF image SR, in which SAIs of each vertical and horizontal views are considered as a sequence, and the long-range geometric dependency is learned via a spatial-angular locally enhanced self-attention layer. Liang et al. [76] proposed a simple yet effective Transformer network (i.e., LFT) for LF image SR. In their method, an angular Transformer is designed to incorporate complementary information among different views, and a spatial Transformer is developed to capture both local and long-range dependencies within each SAI. More recently, Liang et al. [77] investigated the non-local spatial-angular correlations in LF image SR, and developed a Transformer-based network called EPIT to achieve state-of-the-art SR performance. The proposed EPIT achieves a global receptive field along the epipolar line, and is robust to disparity variations. More recently, Jin et al. [78] combined EPIT [77] and DistgSSR [64] to develop a DistgEPIT network for LF image SR. The proposed network achieved state-of-the-art SR accuracy and won the NTIRE 2023 LF image SR Challenge [26].

3. NTIRE 2024 Challenge

In this section, we introduce the NTIRE 2024 LF image SR Challenge. We first introduce the official datasets and toolbox of this challenge. Then, we review the two phases of this challenge. Finally, we summarize the common trends in the submitted solutions.

3.1. Dataset, Toolbox and Evaluation

Training Set. This challenge follows the common settings in [64, 65, 75, 77–79], and uses the EPFL [27], HCInew [28], HCIold [29], INRIA [30] and STFgantry [31] datasets for training. All the 144 LFs in the training set have an angular resolution of 9×9. Challenge participants are required to use these LF images as HR groundtruth to train their models. External training data or models pretrained on other datasets are not allowed in this challenge.

Validation and Test Set. We use the dataset developed in the 1st NTIRE LF image SR challenge for validation and test, as shown in Fig. 1. Both validation and test sets contain 16 synthetic scenes (rendered by 3DS MAX) and 16 real-world scenes (captured by Lytro Illum cameras). Details of the NTIRE-LFSR dataset can be referred to [26].

Table 1. NTIRE 2024 LF Image SR Challenge results, final rankings, and the main characteristics of the solutions. Note that, the average PSNR value achieved on the test set is used for final ranking. The best results are in **red**, the second best results are in **blue**, and the third best results are in **green**.

	Rank	Team	Test Set			Validation Set			#Params	FLOPs	Architec*
			Average	Lytro	Synthetic	Average	Lytro	Synthetic			
Track 1	1	BNU&TMU-AI-TRY★	**30.80/.9332**	**31.00/.9496**	**30.60/.9167**	**32.74/.9508**	**33.46/.9576**	**32.01/.9441**	11.04M	569.30G	Transf
	2	BITSMBU★	**30.73/.9322**	**30.93/.9486**	**30.52/.9159**	**32.64/.9495**	**33.31/.9566**	**31.98/.9425**	5.04M	137.56G	Transf
	3	OpenMeow★	**30.71/.9323**	**30.96/.9491**	**30.46/.9154**	**32.68/.9494**	**33.53/.9577**	**31.82/.9412**	10.63M	353.52G	Transf
	4	IIR-Lab	30.44/.9288	30.96/.9456	30.24/.9120	32.25/.9462	33.03/.9535	31.48/.9388	2.87M	66.13G	Transf
	5	MILab	30.37/.9301	30.58/.9468	30.15/.9134	32.27/.9478	33.03/.9556	31.50/.9399	20.07M	378.92G	Transf
	6	VisionSR	30.31/.9275	30.43/.9439	30.20/.9111	32.38/.9465	33.00/.9530	31.75/.9399	4.52M	153.84G	Transf
	7	Low-level visualist	30.05/.9240	30.19/.9402	29.92/.9079	31.75/.9423	32.68/.9506	30.83/.9341	0.45M	19.33G	CNN
	8	BNU-Small-Potato	29.87/.9226	29.91/.9385	29.82/.9068	31.84/.9427	32.43/.9495	31.25/.9360	25.75M	-	CNN
	9	AQNU-VMIC-team	29.71/.9250	29.53/.9413	29.89/.9087	31.84/.9430	32.44/.9501	31.26/.9360	4.56M	89.16G	CNN
Track 2	1	BITSMBU★	**30.16/.9260**	**30.32/.9425**	**30.00/.9095**	**32.12/.9445**	**32.81/.9518**	**31.43/.9371**	0.66M	19.91G	Transf
	2	Low-level visualist★	**30.05/.9240**	**30.19/.9402**	**29.92/.9079**	**31.75/.9423**	**32.68/.9506**	**30.83/.9341**	0.45M	19.33G	CNN
	3	IIR-Lab★	**29.96/.9238**	**30.14/.9407**	**29.78/.9070**	**31.72/.9411**	**32.47/.9489**	**30.96/.9332**	0.83M	19.47G	Transf
Baselines	-	DistgEPIT [78]	30.66/.9314	30.82/.9475	30.51/.9152	32.71/.9496	33.36/.9562	32.07/.9430	20.34M	566.48G	Hybrid
	-	EPIT [77]	29.87/.9259	29.72/.9420	30.03/.9097	32.04/.9447	32.54/.9507	31.53/.9387	1.47M	76.39G	Transf
	-	DistgSSR [64]	29.64/.9244	29.39/.9403	29.88/.9084	31.75/.9424	32.26/.9490	31.23/.9357	3.58M	65.27G	CNN
	-	Bicubic	25.79/.8378	25.11/.8404	26.46/.8352	27.51/.8714	27.49/.8719	27.53/.8710	-	-	-

Note: "Transf" denotes that the model adopts Transformer as a basic component, "CNN" denotes that the model was developed based on convolutions only.

Note that, all the LF images in the validation and test set are bicubicly downsampled by a factor of 4, and only the LR versions are released to the participants. Challenge participants are required to apply their developed models to the LR LF images, and submit the super-resolved LF images to the CodaLab server for validation and ranking.

Toolbox. We provide BasicLFSR, an open-source and easy-to-use toolbox to facilitate participants to quickly get access to LF image SR and develop their own models. The BasicLFSR toolbox is publicly available at https://github.com/ZhengyuLiang24/BasicLFSR.

Evaluation. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are used as metrics for performance evaluation. The implementation details of PSNR and SSIM can be found in the BasicLFSR toolbox. The submitted results are ranked by the average PSNR values on the test set (both real-world and synthetic scenes).

3.2. Tracks

Track 1: Fidelity Only. This track aims to encourage participants to explore the precision upper bound of LF image SR. In this track, the rankings are determined by the average PSNR value on the test set only. DistgSSR [64] is set as the baseline method in this track. The solutions with PSNR values lower than the DistgSSR will not be ranked in the final leaderboard.

Track 2: Fidelity with Efficiency Constraint. In this track, the model size (*i.e.*, number of parameters) is restricted to 1 MB, and the FLOPs is restricted to 20 G (with an input LF of size $5\times5\times32\times32$). The rankings are determined by the average PSNR value on the test set, but the solutions with model size larger than 1M or FLOPs larger than 20G will not be ranked in the final leaderboard. Bicubic interpolation is set as the baseline method in this track. The solutions with PSNR values lower than the bicubic interpolation will not be ranked in the final leaderboard.

3.3. Challenge Phases

Development Phase. The participants can download the validation set and apply their developed models to the LR LF images to generate their SR versions. A validation leaderboard is available during this phase. The participants can compare their scores with the ones achieved by the baseline models or models developed by other participants.

Test phase. The participants are required to apply their models to the released test set, and submit their super-resolved LF images to the test server. The test server is available online during this phase, and will be closed after the test deadline. The participants are asked to submit the SR results, codes, and a fact sheet of their methods before the given deadline.

3.4. Challenge Results

Among the 125 registered participants, 10 and 6 teams have participated the final test phase and submitted their results, codes, and factsheets. In Track 1, the top 9 teams produced PSNR scores higher than the baseline method DistgSSR [64]. For Track 2, 3 of the 6 teams developed models that meet the efficiency requirement (model size$\leq$1M, FLOPs$\leq$20G). Table 1 reports the PSNR and SSIM scores achieved by these methods on both test and validation sets, together with their major details. We briefly describe these

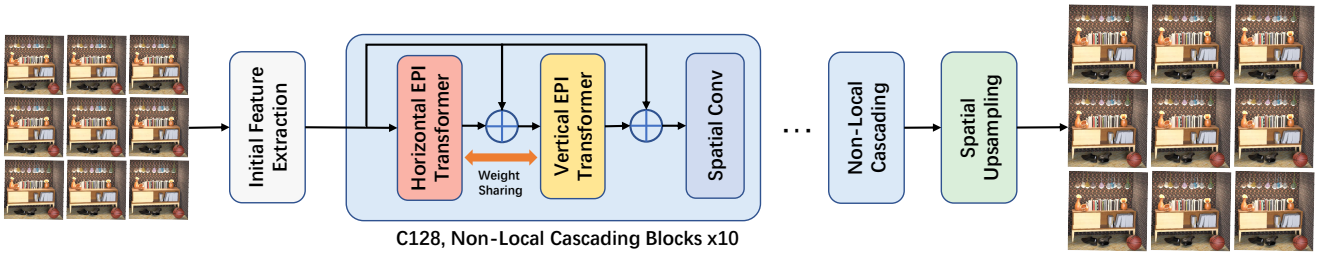

Figure 2. Team BNU&TMU-AI-TRY: The network architecture of the proposed BigEPIT (Track 1).

solutions in Section 4, and introduce the corresponding team members in Appendix 6.

It can be observed from Table 1 that the Track 1 winner BNU&TMU-AI-TRY achieves 0.14 dB improvement in PSNR over DistgEPIT [78] (winner of NTIRE 2023 LF image SR Challenge) on the test set, which pushes LF image SR accuracy to a new height. The winner solution of Track 2, proposed by the BITSMBU team, achieves 30.16 and 32.12 in PSNR on the test and validation set, respectively. It is worth noting that the PSNR scores of all the three solutions in Track 2 surpass those of DistgSSR [64] and EPIT [77] on the test set, which indicates that efficient LF image SR has a large potential.

It is also worth noting that all the proposed methods are based on the deep learning techniques. Most teams adopted Transformers as the basic architecture, while 3 teams build their networks based on CNNs. It seems that Transformers are increasingly popular in LF image SR, and is more powerful in modeling the mapping between LR and HR LF images.

4. Challenge Teams and Methods

4.1. BNU&TMU-AI-TRY: BigEPIT (Track 1★)

The BNU&TMU-AI-TRY team proposed an enhanced network of EPIT [77], called BigEPIT, to handle the disparity problem in LF image SR. They used a Transformer network with repetitive self-attention operations to learn the spatial-angular correlation by modeling the dependencies between each pair of the EPI pixels. Specifically, they increased the channel of feature maps (*i.e.*, $64 \rightarrow 128$) and the number of cascading blocks (*i.e.*, $5 \rightarrow 10$) to improve the model capability. They used the resampling method, which manually specifies different sampling intervals when generating the training data rather than just using the view of the central region.

Data Augmentation: All LFs in the released datasets have an angular resolution of 9×9. In the training stage, they cropped each SAI into patches of size 128×128 with a stride of 32 and used the bicubic downsampling approach to generate LF patches of size 32×32. They performed random horizontal flipping, vertical flipping, and 90-degree rotation to augment the training data. Note that, the spatial and angular dimensions need to be flipped or rotated jointly to maintain LF structures.

Inspired by [78], they also used the augmented data sampling strategy to extract 5×5 SAIs for training and testing, including central, even, and uneven sampling. This strategy explicitly increases the number of images of large disparity LFs, which can improve the robustness of the model to disparity changes.

Regularization: The proposed BigEPIT was trained using the L1 loss and optimized using the Adam method [80] with β_1=0.9, β_2=0.999, and a batch size of 8. Their network was implemented in the framework PyTorch-based *BasicLFSR* on a cluster with four NVIDIA A100 GPUs. The learning rate was initially set to 2×10^{-4} and decreased by a factor of 0.5 for every 15 epochs. The training was stopped after 31 epochs, they selected the best model according to the performance on the validation set.

Ensemble Strategy: They used full-size images as input if GPU memory was available. They utilized 8-set spatial self-ensemble strategies [81] to improve the final results. They fed augmented input LFs independently to the network, including horizontal flip, vertical flip, and rotation, and use the average outputs as predictions. Besides data self-ensemble, they also used a multi-model ensemble strategy to further improve the result, including BigEPIT, DistgEPIT_d_w [78] and RR-HLFSR [82]. DistgEPIT_d_w and RR-HLFSR were trained with augmented data sampling strategy from scratch. However, the multi-model ensemble is time-consuming and introduces little improvements.

4.2. BITSMBU: TriFormer (Track 1★, 2★)

This team participated in two tracks and proposed two networks, respectively.

Track 1: This team proposed a TriFormer network, as shown in Fig. 3(a). TriFormer comprises three parts including initial feature extraction, deep spatial-angular feature learning, and reconstruction. The design of first and last parts follows prior works LFT [76] and EPIT [77]. For the deep spatial-angular feature learning part, this team first employed the proposed Trident block (Triblock) to extract angular and spatial features alternately for three times. Af-

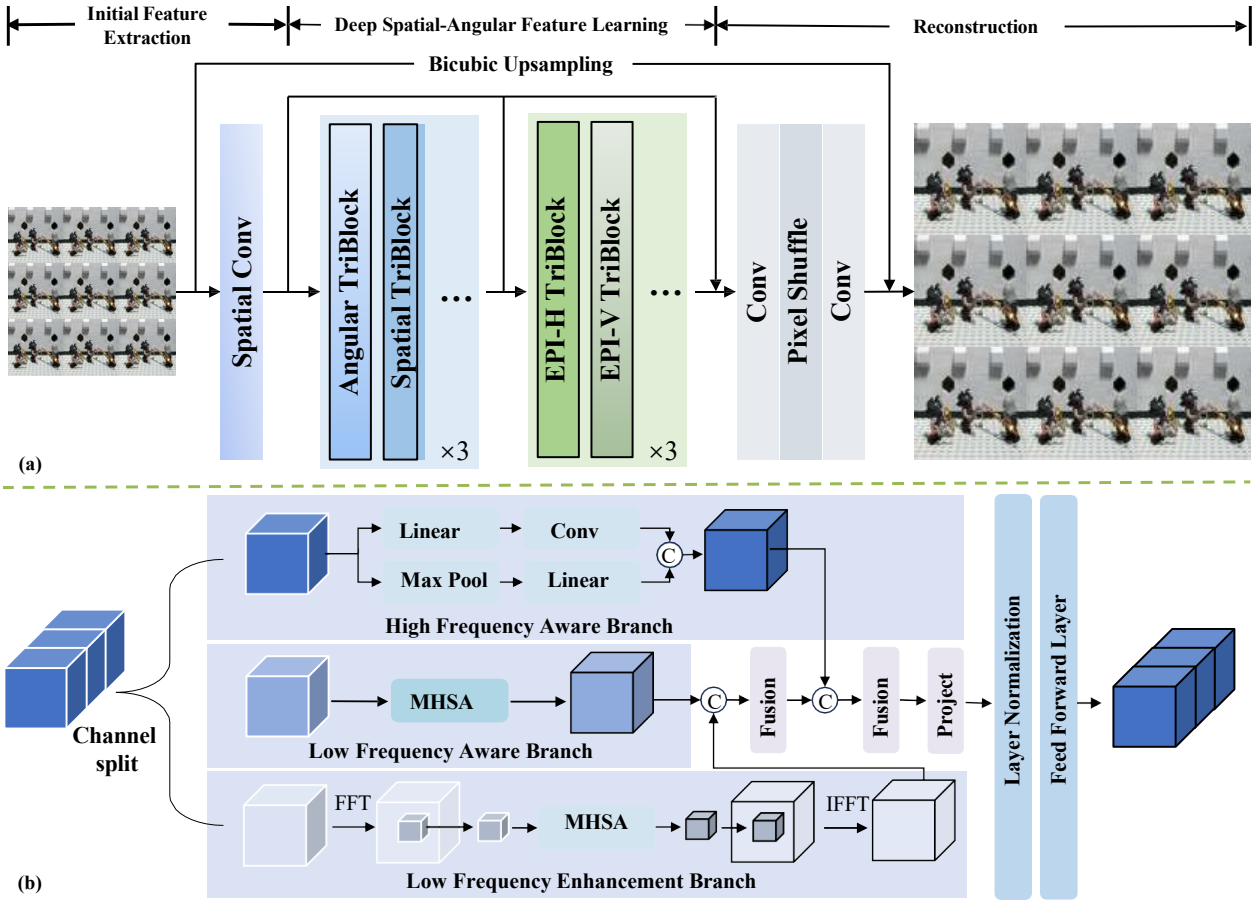

Figure 3. Team BITSMBU: (a) The network architecture of the proposed TriFormer (Track 1); (b) Illustration of the proposed Trident block.

ter that, they employ Triblock to extract spatial-angular correlation on horizontal and vertical EPIs alternately for another three times. Triformer can capture both low-frequency and high-frequency details, and local-global information on the other hand in spatial, angular, and EPI domains.

Trident Block (Triblock): The proposed Triblock consists of three parallel branches: the High-Frequency Aware branch, the Low-Frequency Aware branch, and the Low-Frequency Enhancement branch. For the High-Frequency Aware branch, they adopt convolution to extract high-frequency local details. Meanwhile, they employ a max-pooling layer to preserve the global structural information as compensation for the locality of convolution. Linear layers are used for channel compression. For the Low-Frequency Aware branch, a vanilla multi-head self-attention (MHSA) is applied to capture the long-range dependency of specific LF domains. Lastly, for the Low-Frequency Enhancement branch, they first convert the feature to the frequency domain by performing a Fast Fourier Transform (FFT). Then, the central low-frequency part of the feature is extracted for further enhancement via MHSA. Different from the Low-Frequency Aware branch, they put spatial and angular dimensions all together to excavate the spatial-angular correlations across all views. After three specialized feature extraction processes, the extracted features are fused through convolutions, and the output is further transferred to a feed-forward layer to produce the enhanced feature.

Inference: During inference, they performed the Position-Sensitive Windowing operation proposed by DistgEPIT [78] to preserve the parallax structure of the border region when cropping the full LF image into patches. They also adopted Test Time Augmentation (TTA) to improve the reconstruction quality.

Track 2: Similar to their approach in Track 1, this team also employed the TriFormer network in the Track 2 Fidelity and Efficiency, as shown in Fig. 4. Specifically, the Spatial-Angular Feature Learning module was predominantly implemented using the proposed TriBlocks in four forms: the angular Trident Block (AngTriBlock), spatial

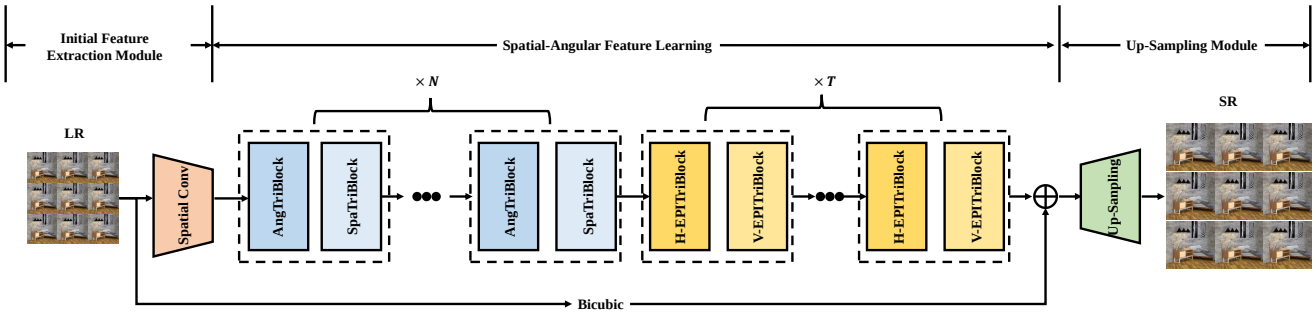

Figure 4. Team BITSMBU: The network architecture of the proposed TriFormer (Track 2).

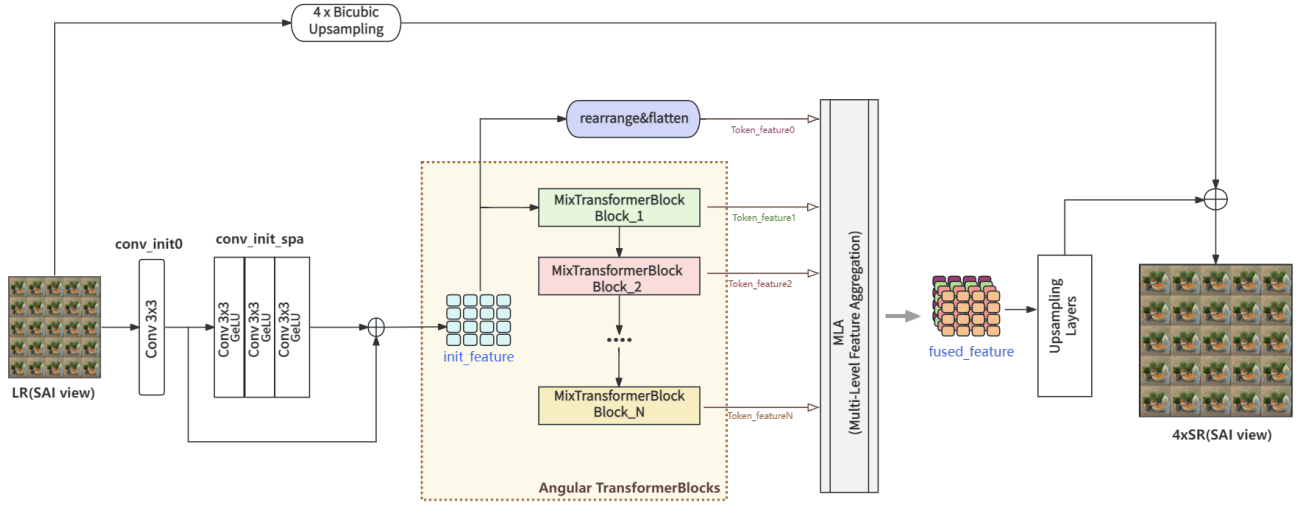

Figure 5. Team OpenMeow: (a) The network architecture of the proposed Fidelity-LF-DET (Track 1).

Trident Block (SpaTriBlock), horizontal EPI Trident Block (H-EPITriBlock), and vertical EPI Trident Block (V-EPI TriBlock). The AngTriBlock and SpaTriBlock were alternately used N times, while the H-EPITriBlock and V-EPI TriBlock were alternately used T times.

4.3. OpenMeow: Fidelity-LF-DET (Track 1★)

This team chose LF-DET [79] as their baseline model which copes both spatial and angular Transformers to capture LF image details and model different disparities. Beyond that, given the scalability and flexibility of LF-DET architecture, they expanded the Transformer model size with more feature channels to discover the potential of large Transformer model, and unify the activation functions in convolutional layers from LeakyReLU to GELU. Referring to Fig. 5, the input SAI obtains spatial features through several convolutional layers, extracting local spatial features. Subsequently, these initial features are passed through multiple Spatial-Angular Global Feature Extraction Transformer blocks for global angular modeling. In the context of these blocks, the output from the preceding block serves as the input to the subsequent block. This sequential process enables hierarchical features to express a wider array of diversified information. In the following learned expressive representations from different blocks will fuse and aggregate via the Feature Aggregation module. Final fused features will be up-sampled through upsampling layers and combined with bicubic up-sampled SR result to obtain the final super-resolved LF image result.

PSW++: Position-Sensitive Windowing Strategy In this team's previous work [78], they introduced an overlapping Position-sensitive windowing (PSW) method as post-processing procedures for large-scale LF image inference. The PSW method utilizes a sliding window approach to crop the image while preserving parallax constraints, particularly around the image border positions. Building upon this work, this team delved deeper into the windowing procedure and proposed an upgraded non-overlapping method called PSW++, which slides windows to collect chops without overlapping the center area of each chop. Given that the

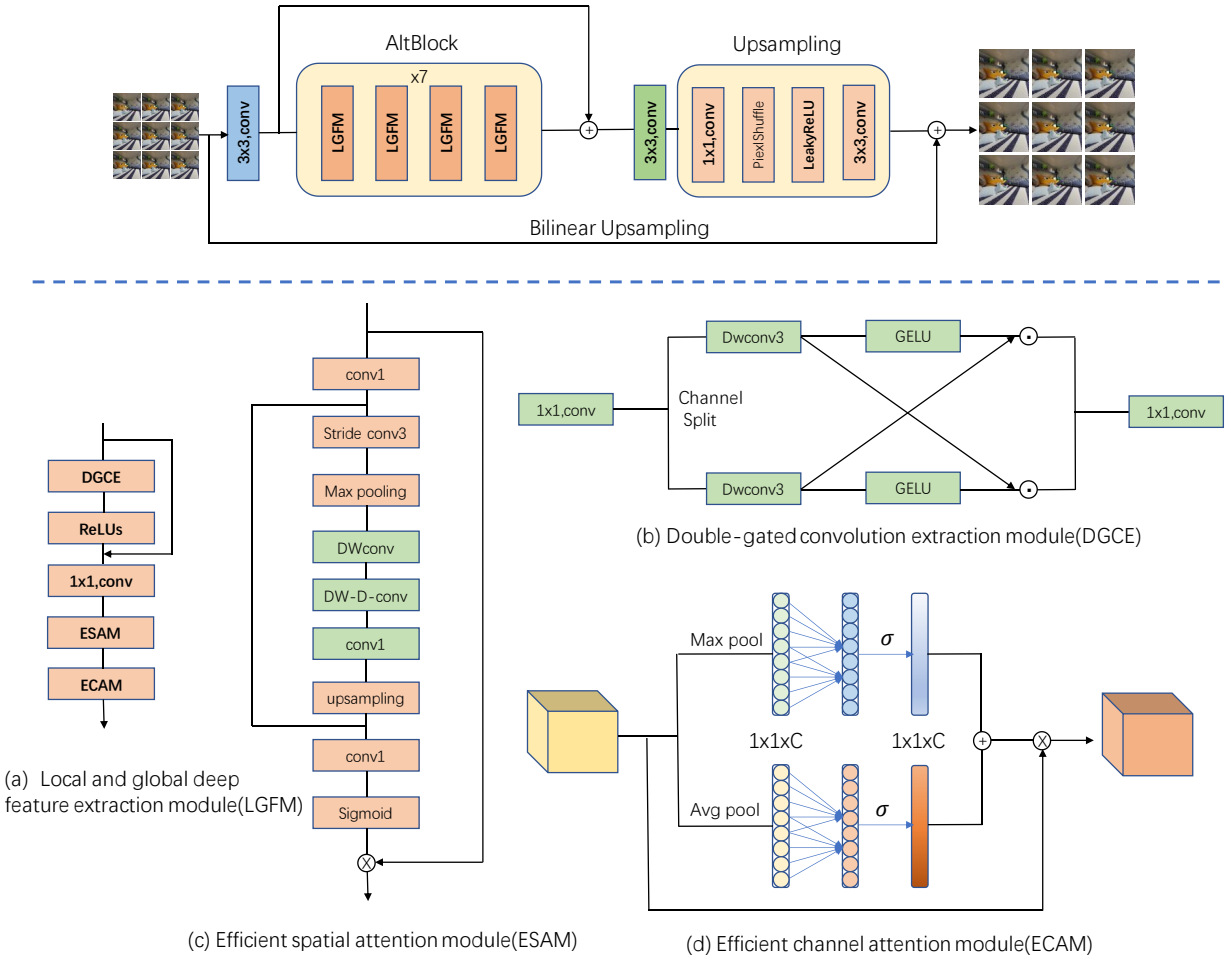

Figure 6. Team Low-level visualist: The network architecture of the proposed LGFN (Track 1, 2).

disparity values gradually decrease from the outermost regions to the center, PSW++ ensures consistency in the disparity structures within the SAI subspace. This avoids the creation of inaccurate disparity structures, ultimately leading to promising performance enhancements as observed through the implementation of PSW++.

Training Strategy: In the Fidelity track, this team would like to fully exploit the capability of large Transformer architecture. To achieve this, they extended the network channels to 160 and trained their Fidelity-LF-DET with a WarmupMultiStepLR learning rate warmup scheduler. This involved a gradual increase of the initial learning rate to 2.5×10^{-4}, followed by reductions of 0.5 via a decay interval set at 15 epochs. They utilized the L1 loss function for optimization and trained the model using the Adam optimizer. The training was carried out with a batch size of 3 and β_1 and β_2 values set to 0.99 and 0.999, respectively. The fidelity-LF-DET model was implemented in PyTorch and the training was conducted on a system equipped with three NVidia V100 GPUs.

4.4. Low-level visualist: LGFN (Track 1, 2★)

The Low-level visualist team participated in two tracks with the proposed LGFN method. As shown in Fig. 6, the proposed pipeline includes the shallow feature extraction, deep feature extraction $H_{LGFM}(.)$, and up-sampling modules. The LGFM consists of three parts including the double-gated convolution extraction module (DGCE), the efficient spatial attention module (ESAM), and the efficient channel attention module (ECAM). Based on the similarity of LF sub-aperture images, LGFN achieves LF image SR by learning local and global features. Specifically, they design a lightweight convolution module to extract the local features of the LF image by modulation. In addition, in order to learn the global features, they design an efficient spatial attention module and an efficient channel attention module by enlarging the receptive field through large kernel convolution.

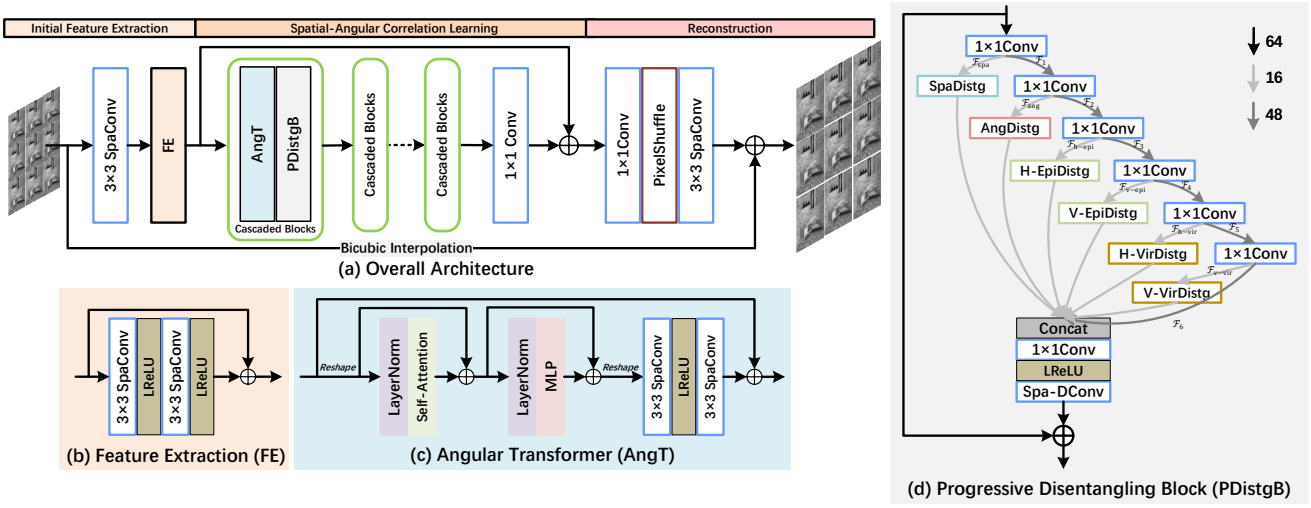

Figure 7. Team IIR-Lab: The network architecture of the proposed PDistgNet (Track 1, 2).

The proposed FGFN was trained using the L1 loss and FFT Charbonnier loss with weights of 0.01 and 1, respectively. The model was implemented in PyTorch on a PC with an NVidia RTX 3060 GPU. The learning rate was initially set to 2×10^{-4} and decreased by a factor of 0.5 for every 15 epochs. The training was stopped after 100 epochs.

4.5. IIR-Lab: PDistgNet (Track 1, 2★)

This team participated in two tracks with the proposed PDistgNet and its large version PDistgNet-L.

Track 2: Inspired by IMDN [83] and disentangling mechanism [84, 85], the IIR-Lab team introduced a progressive disentangling block (PDistgB), which progressively disentangles the LF feature into multiple subspaces for leveraging the structure priors of LF and reduce the computational costs. Additionally, considering the macro-pixel pattern has a relatively small size (*i.e.*, 5×5), applying Transformer on the angular domain is also an efficient choice to incorporate the angular correlations. This team developed PDistgNet with PDistgB and angular Transformer (AngT) for efficient LF image SR.

An overview of the proposed PDistgNet is shown in Fig. 7(a). Concretely, their method contains three steps: initial feature extraction, spatial-angular correlation learning, and reconstruction. The initial feature extraction consists of a 1×1 convolution and a feature extraction (FE) module (Fig. 7(b)) to extract the intra-view correlations. After that, they apply N cascaded blocks, which are composed of the angular transformer (AngT) and Progressive disentangling block (PDistgB), to leverage the inherent spatial-angular correlations of LF. The structures of AngT and PDistgB are depicted in Fig. 7(c) and (d), respectively. Specifically, they follow previous work [86] to build their AngT, and further apply two spatial convolutions to enhance the feature representations. In PDistgB, multiple feature splits and disentanglements are performed. In addition to the spatial, angular, and EPI domains, they also follow previous work [85] to conduct horizontal/vertical virtual-slit domain disentangling. Finally, they up-sample the spatial resolution of the LF feature to generate HR LF image. A global residual connection with bicubic interpolation is deployed to feed low-frequency information to the output.

Track 1: They also developed a large version of PDistgNet, called PDistgNet-L, for Track 1. Specifically, N is set to 4 in PDistgNet, and N is set to 16 in PDistgNet-L.

4.6. MILab: HA-DET (Track 1)

The MILab team proposed a HA-DET model which is an upgraded version of LF-DET [79] enhanced by HAT [87] modules. The main module of HA-DET is the Spatial-Angular Separable Transformer Block. In this block, they sequentially conducted spatial and angular transformer encoding dimensions under the representations of SAI and MacPI (see Fig. 8). Considering that the vanilla Transformer block suffers from huge computation and memory consumption incurred by high-resolution SAIs, they deployed the HAT module in the spatial stage, which includes a parallel swin-transformer module and channel attention module. As for the angular stage, they kept the basic structure in LF-DET but expanded the scope of the Transformer to capture more comprehensive angular information.

Ensemble Strategy. They used three different configurations to implement their HA-DETs. Specifically, the three models in terms of channel numbers, heads of MSA, window size, and encoder depths were respectively {96, 4, 8, 8}, {120, 6, 16, 6}, and {120, 6, 16, 8}. The numbers

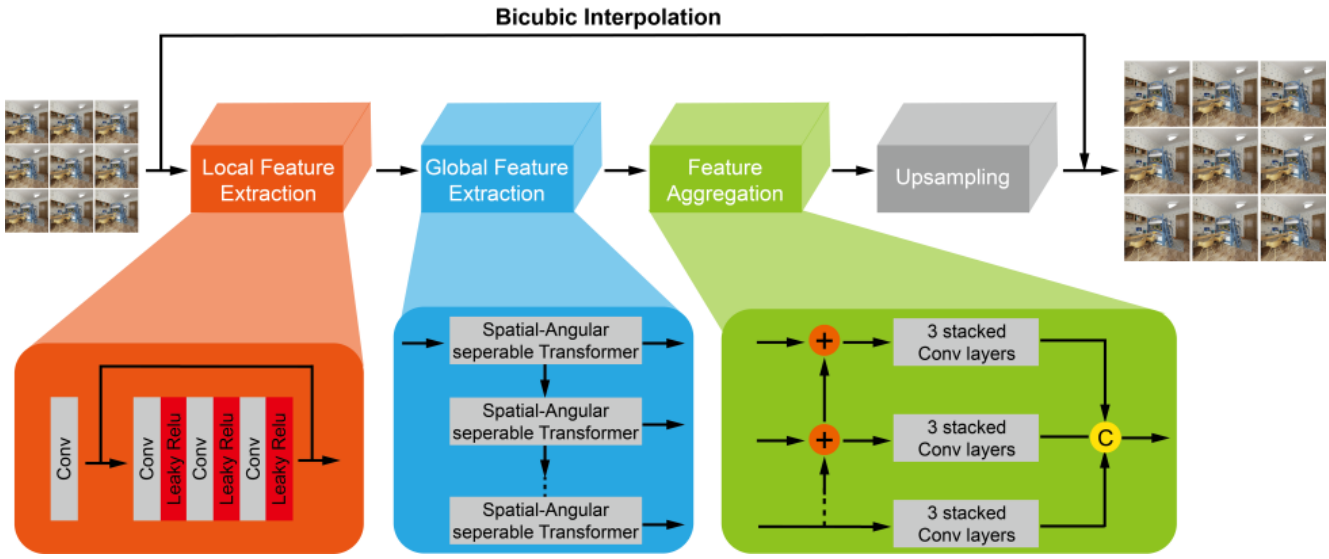

Figure 8. Team MILab: The network architecture of the proposed HA-DET (Track 1).

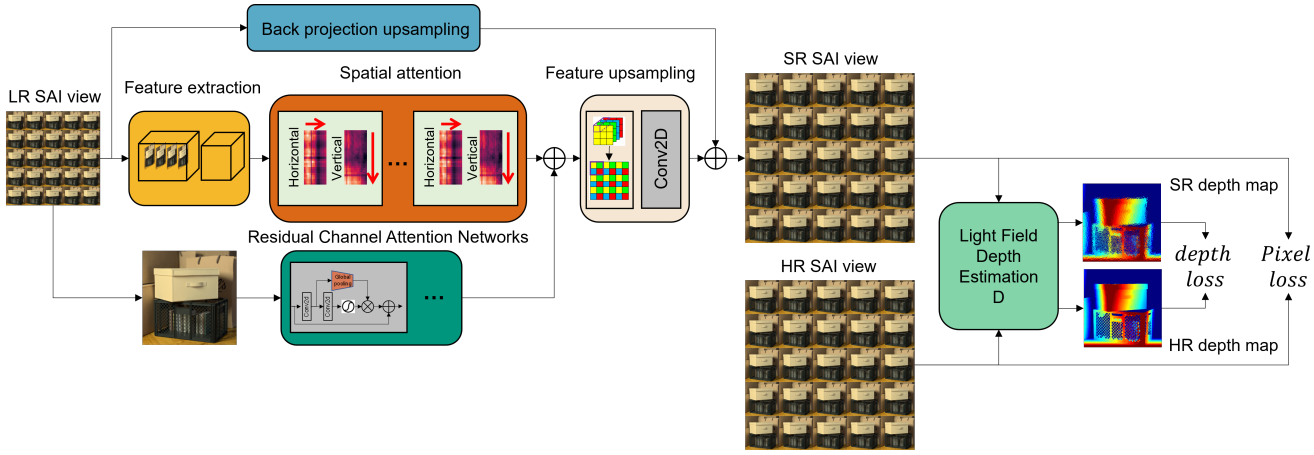

Figure 9. Team VisionSR: The network architecture of the proposed E-EPIT (Track 1).

of the spatial and angular Transformers were both set to 4 in the three models. The numbers of parameters of these models were 12.9M, 15.1M, and 20.0M, respectively. In the test phase, they utilized the self-ensemble approach in EDSR [88] to further enhance the model performance.

4.7. VisionSR: E-EPIT (Track 1)

Inspired by EPIT [77], the VisionSR team proposed an enhanced EPIT (E-EPIT) network to improve data fidelity. The overall structure is shown in Fig. 9. The Given input LR images are processed through two separate pathways: the SAI path and the MacPI path. The SAI path is to use LR SAI views as inputs to go through the 3D convolution process (Feature extraction in the yellow box) and obtain the joint spatial-angular feature representation. The spatial attention (the orange box in the figure) is adopted from EPIT [77]. The idea is to convert the feature horizontally and vertically and use a multi-head attention module to extract the non-local feature correlations.

The MacPI path is to use pixel shuffling operator to rearrange the LR SAI images into MacPI patterns. The MacPI image is then processed by the Residual Channel Attention Network (RCAN), which contains multiple residual blocks to enhance the global feature representation. Both results from MacPI and SAI are combined to learn the residues between LR and HR images. To match the dimension of the target image, they use pixel shuffle again to fuse the features for the final image. Meanwhile, this team used back projection to initially upsample the LR image and use it to add back to the final output. The idea of back projection is to iteratively update the residues between LR and SR images to reduce pixel distortion.

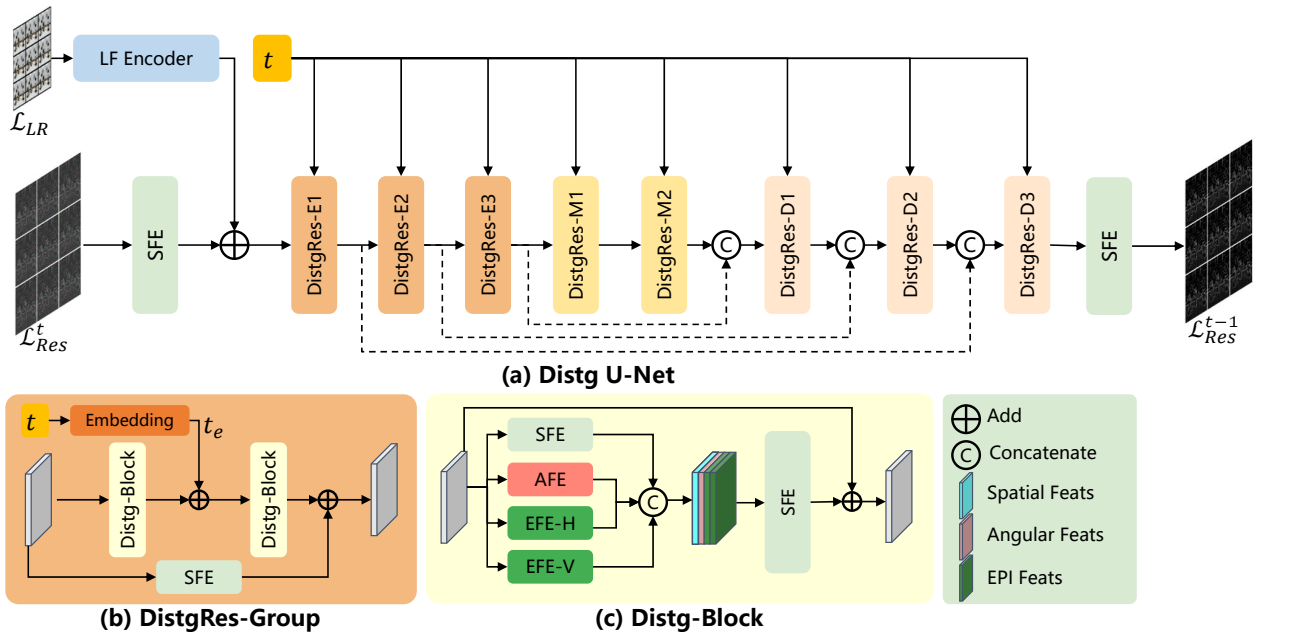

Figure 10. Team BNU-Small-Potato: The network architecture of the proposed LFSRDiff (Track 1).

Loss Function. This team used the L1 loss to minimize the pixel distortion. To facilitate model training, they additionally utilized a pre-trained LF Depth estimation module to enhance overall optimization. To extract the depth, they used pre-trained OACCNet [4]. Since the pre-trained model only accepts inputs with an angular resolution of 9×9, they designed the following process for loss calculation: 1) they prepared the ground truth 9×9 SAI images as $Y_{9\times9}$, 2) then used $Y_{9\times9}$ as inputs to OACCNet to calculate the disparity mask M and corresponding depth Z, 3) they replaced the center 5×5 views of $Y_{9\times9}$ by the SR images and obtain estimated 9×9 SAI images as $Y'_{9\times9}$, and 4) they used the disparity mask M and estimated 9×9 SAI images to obtain the depth map Z'. Finally, They calculated the L1 loss between Z and Z' for supervision.

4.8. BNU-Small-Potato: LFSRDiff (Track 1)

The BNU-Small-Potato team incorporated the LF disentanglement mechanism to the LFSRDiff [89], the diffusion-based LF image SR model. The main components of the proposed Distg U-Net contain three DitsgRes-Groups in the encoder (DistgRes-E), two DitsgRes-Groups in the bottleneck (DistgRes-M), and three DitsgRes-Groups in the decoder (DistgRes-D). Skip connections are used between the encoder and the decoder at the same level. Each DitsgRes-Group (see Fig.10(b)) consists of a residual spatial convolution, two disentangling blocks (Distg-Blocks). The timestep embedding t_e is added to the extracted feature of the first Distg-Blocks through spatial replication. DistgRes-E and DistgRes-D have an extra downsampling or upsampling operation.

4.9. AQNU-VMIC-team: MRVRNet (Track 1)

This team designed a Multi-Representation View Reconstruction Network (MRVRNet) to fully explore the structural properties of the LF by utilizing the properties of different representations. They followed DistgSSR [64] in the MacPI Branch, which organizes the input LF into a MacPI and refines the process by leveraging disentanglement blocks and channel attention to effectively extract features. In the SAI Branch, they organized the input LF into a sub-aperture array and perceive LF texture by extracting gradient information and using Resblock to effectively mitigate the loss of texture and structural details, especially for high-texture areas. Channel attention and angular attention [90] are also introduced to enhance feature representation and further explore angular consistency.

5. Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 61921001, U20A20185, 61972435), and the Outstanding Youth Foundation in Hunan Province (No. 2024JJ2063). This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2024 sponsors: Meta Reality Labs, OPPO, KuaiShou, Huawei and University of Würzburg (Computer Vision Lab).

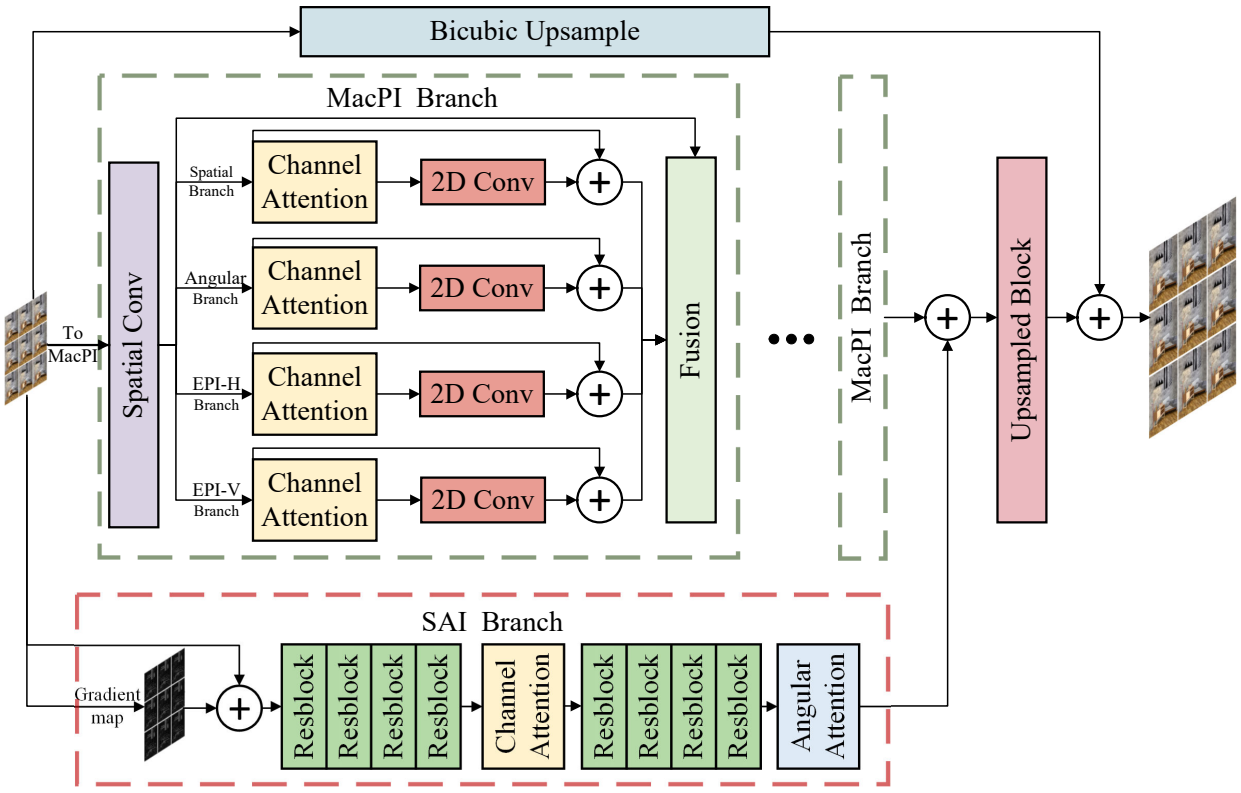

Figure 11. Team AQNU-VMIC-team: The network architecture of the proposed MRVRNet (Track 1).

6. Teams and Affiliations

Challenge Organizers

Members:
Yingqian Wang[1] (wangyingqian16@nudt.edu.cn),
Zhengyu Liang[1] (zyliang@nudt.edu.cn),
Qianyu Chen[1] (chenqianyu18@nudt.edu.cn),
Longguang Wang[2] (wanglongguang15@nudt.edu.cn),
Jungang Yang[1] (yangjungang@nudt.edu.cn),
Radu Timofte[3] (radu.timofte@uni-wuerzburg.de),
Yulan Guo[1] (yulan.guo@nudt.edu.cn).

Affiliations:
[1]National University of Defense Technology
[2]Aviation University of Air Force
[3]Computer Vision Lab, University of Würzburg

(1) The BNU&TMU-AI-TRY Team - Track 1★

Members:
Wentao Chao[1](chaowentao@mail.bnu.edu.cn), Yiming Kan[1], Xuechun Wang[1], Fuqing Duan[1], Guanghui Wang[2]

Affiliations:
[1]Beijing Normal University
[2]Toronto Metropolitan University

(2) The BITSMBU Team - Track 1★, 2★

Members:
Wang Xia[1,2](3220221027@bit.edu.cn), Ziqi Wang[1,2], Yue Yan[1,2], Peiqi Xia[1,2], Shunzhou Wang[3] , Yao Lu[1,2]

Affiliations:
[1]Beijing Institute of Technology
[2]Shenzhen MSU-BIT University
[3]Peking University Shenzhen Graduate School

(3) The OpenMeow Team - Track 1★

Members:
Angulia Yang[1](yangying.angulia@bigo.sg), Kai Jin[1], Zeqiang Wei[3], Sha Guo[2] Mingzhi Gao[1] Xiuzhuang Zhou[3]

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

(4) The Low-level visualist Team - Track1, 2★

Members:
Zhongxin Yu[1] (wuyizhizi555@163.com), Shaofei Luo[1], Cheng Zhong[1], Shaorui Chen[1], Long Peng[2], Yuhong He[3]

Affiliations:
[1]Fujian Normal University
[2]University of Science and Technology of China
[3]Northeastern University

(5) The IIR-Lab Team - Track1, 2★

Members:
Gaosheng Liu[1] (gaoshengliu@tju.edu.cn), Huanjing Yue[1], Jingyu Yang[1]

Affiliations:
[1]School of Electrical and Information Engineering, Tianjin University

(6) The MILab Team - Track1

Members:
Zhengjian Yao[1] (zj.yao@stu.pku.edu.cn), Jiakui Hu[1], Lujia Jin[1]

Affiliations:
[1]Peking University

(7) The VisionSR Team - Track1

Members:
Zhi-Song Liu[1] (zhisong.liu@lut.fi), Chenhang He[2], Jun Xiao[2], Xiuyuan Wang[2]

Affiliations:
[1]Lappeenranta–Lahti University of Technology LUT
[2]The Hong Kong Polytechnic University

(8) The BNU-Small-Potato Team - Track1

Members:
Zonglin Tian[1] (zonglintian@mail.bnu.edu.cn)

Affiliations:
[1]Beijing Normal University

(9) The AQNU-VMIC-team Team - Track1

Members:
Yifan Mao[1] (maoyifan1998@sina.comn), Deyang Liu[1], Shizheng Li[1], Ping An[2]

Affiliations:
[1]Anqing Normal University
[2]Shanghai University

References

[1] Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy. Using plane+ parallax for calibrating dense camera arrays. In *CVPR*, 2004. 1

[2] 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, 2018. 1

[3] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4748–4757, 2018. 1

[4] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 11

[5] Wentao Chao, Xuechun Wang, Yingqian Wang, Liang Chang, and Fuqing Duan. Learning sub-pixel disparity distribution for light field depth estimation. *arXiv preprint arXiv:2208.09688*, 2022. 1

[6] Ryan S Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr, and Paul Debevec. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Transactions on Graphics*, 37(6):1–15, 2018. 1

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

[8] Gaochang Wu, Yebin Liu, Lu Fang, and Tianyou Chai. Revisiting light field rendering with deep anti-aliasing neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[9] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. 1

[10] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[11] Benjamin Attal, Jia-Bin Huang, Michael Zollhoefer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[12] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074, 2019. 1

[13] Tao Dai, Mengxi Ya, Jinmin Li, Xinyi Zhang, Shu-Tao Xia, and Zexuan Zhu. Cfgn: A lightweight context feature guided network for image super-resolution. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023. 1

[14] Jinmin Li, Tao Dai, Mingyan Zhu, Bin Chen, Zhi Wang, and Shu-Tao Xia. Fsr: A general frequency-oriented framework to accelerate image super-resolution networks. In *AAAI*, volume 37, pages 1343–1350, 2023. 1

[15] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2021. 1

[16] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *CVPR*, pages 4917–4926, 2021. 1

[17] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, pages 517–532, 2018. 1

[18] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *CVPRW*, pages 457–466, 2022. 1, 3

[19] Juncheng Li, Zehua Pei, Wenjie Li, Guangwei Gao, Longguang Wang, Yingqian Wang, and Tieyong Zeng. A systematic survey of deep learning-based single-image super-resolution. *ACM Computing Surveys*, 2024. 1

[20] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[21] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020. 1

[22] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *CVPRW*, pages 766–775, 2021. 1

[23] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. In *ACM MM*, pages 1985–1993, 2021. 1

[24] Xinyi Ying, Longguang Wang, Yingqian Wang, Weidong Sheng, Wei An, and Yulan Guo. Deformable 3d convolution for video super-resolution. *IEEE Signal Processing Letters*, 27:1500–1504. 1

[25] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using HR optical flow estimation. *IEEE Transactions on Image Processing*, 2020. 1

[26] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, Kai Jin, Zeqiang Wei, Angulia Yang, Sha Guo, et al. Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results. In *CVPRW*, pages 1320–1335, 2023. 1, 3

[27] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 1, 3

[28] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision (ACCV)*, pages 19–34, 2016. 1, 3

[29] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization (VMV)*, volume 13, pages 225–226, 2013. 1, 3

[30] 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. 1, 3

[31] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7), 2008. 1, 3

[32] Cosmin Ancuti, Codruta O Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2024 dense and non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[33] Nikola Banić, Egor Ershov, Artyom Panshin, Oleg Karasev, Sergey Korchagin, Shepelev Lev, Alexandr Startsev, Daniil Vladimirov, Ekaterina Zaychenkova, Dmitrii R Iarchuk, Maria Efimova, Radu Timofte, Arseniy Terekhin, et al. NTIRE 2024 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[34] Ren Yang, Radu Timofte, et al. NTIRE 2024 challenge on blind enhancement of compressed image: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[35] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei WU, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[36] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, et al. The ninth NTIRE 2024 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[37] Zheng Chen, Zongwei WU, Eduard Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. NTIRE 2024 challenge on image super-resolution (×4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[39] Longguang Wang, Yulan Guo, Juncheng Li, Hongda Liu, Yang Zhao, Yingqian Wang, Zhi Jin, Shuhang Gu, Radu Timofte, et al. NTIRE 2024 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[40] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, et al. NTIRE 2024 challenge on HR depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[41] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Wangmeng Zuo, Radu Timofte, et al. NTIRE 2024 challenge on bracketing image restoration and enhancement: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[42] Nicolas Chahine, Marcos V. Conde, Sira Ferradans, Radu Timofte, et al. Deep portrait quality assessment. a NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[43] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[44] Jie Liang, Qiaosi Yi, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Radu Timofte, Lei Zhang, et al. NTIRE 2024 restore any image model (RAIM) in the wild challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[45] Marcos V. Conde, Florin-Alexandru Vasluianu, Radu Timofte, et al. Deep raw image super-resolution. a NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[46] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[47] Xiaoning Liu, Zongwei WU, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. NTIRE 2024 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[48] 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. 2

[49] 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. 2

[50] 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. 2

[51] Martin Alain and Aljosa Smolic. Light field denoising by sparse 5d transform domain collaborative filtering. In *International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017. 2

[52] Mattia Rossi and Pascal Frossard. Graph-based light field super-resolution. In *International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017. 2

[53] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 24–32, 2015. 2

[54] 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. 2

[55] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11046–11055, 2019. 2

[56] 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*, 2021. 2

[57] 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*, 2019. 2

[58] Nan Meng, Hayden Kwok-Hay So, Xing Sun, and Edmund Lam. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[59] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2260–2269, 2020. 2

[60] 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*, 2020. 2

[61] Yu Mo, Yingqian Wang, Chao Xiao, Jungang Yang, and Wei An. Dense dual-attention network for light field image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4431–4443, 2021. 2

[62] 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

[63] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[64] 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, 3, 4, 5, 11

[65] 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. 3

[66] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10010–10019, 2021. 3

[67] 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. 3

[68] Zeyu Xiao, Yutong Liu, Ruisheng Gao, and Zhiwei Xiong. Cutmib: Boosting light field super-resolution via multi-view image blending. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[70] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 3

[71] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 3

[72] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 3

[73] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 3

[74] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 3

[75] 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 (AAAI)*, 2022. 3

[76] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. Light field image super-resolution with transformers. *IEEE Signal Processing Letters*, 2022. 3, 5

[77] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, Zhou Shilin, and Yulan Guo. Learning non-local spatial-angular correlation for light field image super-resolution. *arXiv preprint arXiv:2302.08058*, 2023. 3, 4, 5, 10

[78] Kai Jin, Angulia Yang, Zeqiang Wei, Sha Guo, Mingzhi Gao, and Xiuzhuang Zhou. Distgepit: Enhanced disparity learning for light field image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 3, 4, 5, 6, 7

[79] Ruixuan Cong, Hao Sheng, Da Yang, Zhenglong Cui, and Rongshan Chen. Exploiting spatial and angular correlations with deep efficient transformers for light field image super resolution. *IEEE Transactions on Multimedia*, 2023. 3, 7, 9

[80] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning and Representation (ICLR)*, 2015. 5

[81] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1865–1873, 2016. 5

[82] 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. 5

[83] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019. 9

[84] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022. 9

[85] Manchang Jin, Gaosheng Liu, Kunshu Hu, Xin Luo, Kun Li, and Jingyu Yang. Physics-informed ensemble representation for light-field image super-resolution. *arXiv preprint arXiv:2305.20006*, 2023. 9

[86] 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. 9

[87] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 9

[88] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144, 2017. 10

[89] 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. 11

[90] Gaosheng Liu, Huanjing Yue, Kun Li, and Jingyu Yang. Disparity-guided light field image super-resolution via feature modulation and recalibration. *IEEE Transactions on Broadcasting*, 69(3):740–752, 2023. 11