

Toward Real-World Light Field Super-Resolution

Zeyu Xiao* Ruisheng Gao* Yutong Liu Yueyi Zhang Zhiwei Xiong†

University of Science and Technology of China

{zeyuxiao, grsmc4180, ustclyt}@mail.ustc.edu.cn {zhyuey, zwxiong}@ustc.edu.cn

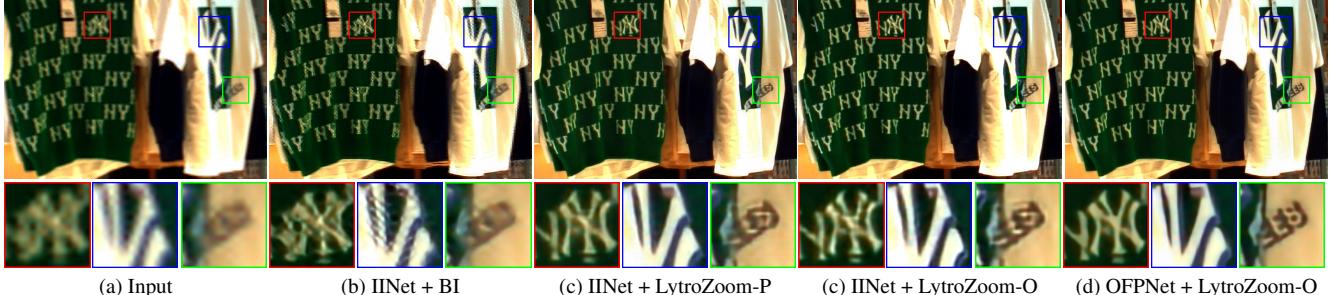


Figure 1. Examples of $\times 4$ real-world light field SR. We collect the first real-world paired LR-HR light field SR dataset (LytroZoom) with authentic degradations to address the real-world light field SR task. We show the super-resolved results of existing advanced light field SR method and our proposed OFPNet. We show (a) a real-world light field central view image captured by a Lytro ILLUM camera at a focal length of 60 mm, SR results generated by (b) IIINet [42] trained on the mixed bicubic downsampled synthetic dataset, (c) IIINet trained on LytroZoom-P, (d) IIINet fine-tuned on LytroZoom-O, and (e) the proposed OFPNet fine-tuned on LytroZoom-O.

Abstract

Deep learning has opened up new possibilities for light field super-resolution (SR), but existing methods trained on synthetic datasets with simple degradations (e.g., bicubic downsampling) suffer from poor performance when applied to complex real-world scenarios. To address this problem, we introduce LytroZoom, the first real-world light field SR dataset capturing paired low- and high-resolution light fields of diverse indoor and outdoor scenes using a Lytro ILLUM camera. Additionally, we propose the Omni-Frequency Projection Network (OFPNet), which decomposes the omni-frequency components and iteratively enhances them through frequency projection operations to address spatially variant degradation processes present in all frequency components. Experiments demonstrate that models trained on LytroZoom outperform those trained on synthetic datasets and are generalizable to diverse content and devices. Quantitative and qualitative evaluations verify the superiority of OFPNet. We believe this work will inspire future research in real-world light field SR. Code and dataset are available at <https://github.com/zeyuxiao1997/RealLFSR>.

1. Introduction

The light field imaging technique enables the capture of the light rays not only at different locations but also from different directions [4]. The limited spatial resolution caused by the essential spatial-angular trade-off restricts the capability of light field in practical applications, such as post-capture refocusing [47, 66], disparity estimation [60, 63, 80], and seeing through occlusions [30, 65, 79]. Attempting to recover a high-resolution (HR) light field from its low-resolution (LR) observation, light field super-resolution (SR) has emerged as a significant task in both academia and industry. Benefiting from the rapid development of deep learning techniques, convolutional neural network (CNN) based and vision Transformers based methods have demonstrated promising performance for light field SR [10, 19, 29, 39, 40, 42, 46, 58, 62–64, 67, 70, 72–74, 77, 78]. They outperform traditional non-learning-based methods [2, 10, 52] with appreciable improvements.

However, existing deep methods remain limited because they are trained on simulated light field datasets which assume simple and uniform degradations (e.g., bicubic down-sampling) due to the natural difficulty of collecting LR-HR light field pairs. Degradations in real applications are much more complicated and such simulated degradations usually deviate from real ones. This degradation mismatch makes

* Both authors contribute equally to this work.

† Corresponding author.

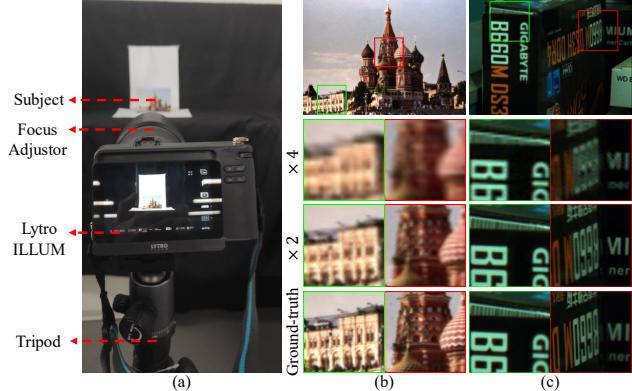


Figure 2. (a) The capturing system of the LytroZoom dataset. (b) One example of pixel-wise aligned pairs from LytroZoom-P. (c) One example of pixel-wise aligned pairs from LytroZoom-O.

existing light field SR methods unpractical in real-world scenarios [8, 9]. Fig. 1 shows the light field SR results of a real-world light field captured by a Lytro ILLUM camera. We utilize the advanced IINet [63] to train several light field SR models using the mixed simulated dataset with bicubic degradation (IINet + BI) and light field pairs with authentic distortions in our LytroZoom (IINet + LytroZoom-P/LytroZoom-O). The results clearly show that, the IINet trained on the simulated dataset (Fig. 1(b)) is less effective in super-resolving on a real-world light field (*e.g.*, a light field captured by Lytro ILLUM).

To remedy the above mentioned problem, it is highly desired that we have a light field SR dataset of paired LR-HR pairs more consistent with real-world degradations. However, collecting such a real-world light field SR dataset is non-trivial since the ground-truth HR light fields are tough to obtain. Inspired by [6], in which a real-world single image SR dataset is built upon the intrinsic resolution and field-of-view (FoV) degradation in realistic imaging systems, we capture images of the same scene (*i.e.*, indoor and outdoor scenes) using a Lytro ILLUM camera (Fig. 2(a)) with different adjusted focal lengths. LR and HR light field pairs at different scales (*e.g.*, $\times 2$ and $\times 4$) can be collected by adjusting the focal length. We utilize the alignment algorithm proposed in [5] to rectify distortions caused by different FoVs, including spatial misalignment, intensity variation, and color mismatching. Scenes that could not be rectified are eliminated from the dataset. As a result, we collect LytroZoom, a dataset of 94 aligned pairs featuring city scenes printed on postcards (LytroZoom-P), as well as 63 aligned pairs captured with outdoor scenes (LytroZoom-O), as illustrated in Fig. 2(b) and Fig. 2(c). Featuring scenes with spatial diversity and fine-grained details printed on postcards, and depth variations in outdoor scenes, LytroZoom presents a benchmark dataset for light field SR algorithms in real-world scenarios. As can be seen in Fig. 1(c)-(d), IINet trained on LytroZoom-P and fine-

tuned on LytroZoom-O delivers much better results than the one trained on the simulated data.

Compared with those simulated datasets, the degradation process in LytroZoom is much more complicated. In particular, real-world degradations exists in all frequency components and are spatially variant [15, 16, 49]. This motivates us to design a network which can consider omni-frequency information and enhance cross-frequency representations. In this paper, we propose the Omni-Frequency Projection Network (OFPNet) to efficiently solve the real-world light field SR problem. Specifically, we first decompose the input LR light field into high-frequency, middle-frequency and low-frequency components. Then, in order to improve the corresponding frequency representations, we employ multiple interactive branches that include frequency projection (FP) operations. Cores of the FP operation are iterative upsampling and downsampling layers to learn non-linear relationships between LR and HR frequency components for super-resolving a real-world light field. As shown in Fig. 1(e), our OFPNet can generate better results.

This work makes three key contributions.

(1) We collect LytroZoom, the first real-world paired LR-HR light field SR dataset, which overcomes the limitations of synthetic light field SR datasets and provides a new benchmark for training and evaluating real-world light field SR methods.

(2) We demonstrate that light field SR models trained on LytroZoom perform better on real-world light fields than those trained on synthetic datasets.

(3) We propose OFPNet, a novel baseline network for real-world light field SR, achieving superior results compared to existing methods.

2. Related work

Single image SR. Single image SR methods have traditionally been trained on manually synthesized LR images using pre-defined downsampling kernels, such as bicubic interpolation [11, 14, 17, 26, 31, 34, 35, 38, 41, 43, 45, 48, 54, 82, 83]. However, these methods are not directly applicable to real-world images due to their more complex degradation kernels. Blind SR methods have recently emerged as a solution, which can be categorized into explicit degradation modeling [3, 18, 25, 32, 57, 76], capture training pairs [5, 6, 59, 69, 81] and generate training pairs [27, 44, 75, 84]. Inspired by the success of single image SR, we propose LytroZoom, the first real-world paired LR-HR light field SR dataset, and a novel OFPNet for real-world light field SR. The view-dependent degradations in LytroZoom are implicitly modeled by the OFPNet, providing an effective solution for real-world light field SR.

Light field SR. Traditional non-learning methods depend on geometric [37, 52] and mathematical [2] modeling of the 4D light field structure to super-resolve the reference

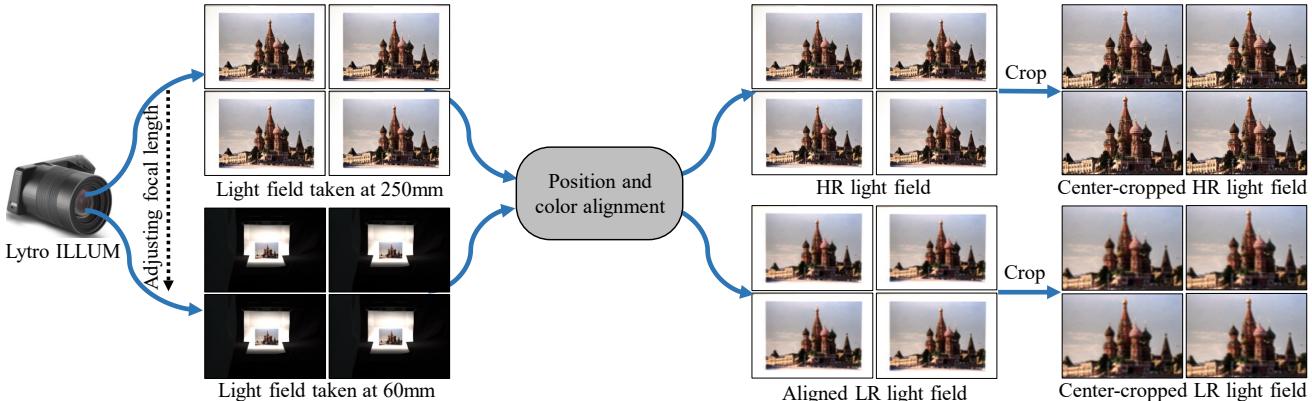


Figure 3. The processing pipeline of LytroZoom. Here we show how to obtain a $\times 4$ LR-HR real-world light field pair from the postcard.

view through projection and optimization techniques. Deep methods now dominate light field SR due to their promising performance. As a pioneering work along this line, Yoon *et al.* [73] propose the first light field SR network LFCNN by reusing the SRCNN [14] architecture with multiple channels. After that, several methods have been designed to exploit across-view redundancy in the light field, either explicitly [10, 29, 62, 78] or implicitly [46, 63, 64, 72, 74]. Transformer-based methods have recently demonstrated the effectiveness in light field SR [39, 40, 58]. Recently, Cheng *et al.* [9] propose a zero-shot learning framework to solve the domain shift problem in light field SR methods. This work can be regarded as a step towards light field SR in real-world scenarios, but real degradations in light field SR are not approached. In this paper, we collect the first real-world paired LR-HR light field SR dataset.

Light field SR datasets. Several popular datasets, including EPFL [51], HCInew [23], HCIold [68], INRIA [33], STFgantry [55], and STFlytro [50] are widely used for training and evaluating light field SR methods. Recently, Sheng *et al.* [53] collect UrbanLF, a comprehensive light field dataset containing complex urban scenes for the task of light field semantic segmentation. This dataset has the potential to be extended to light field SR as well. In all these datasets, the LR light fields are synthesized by the bicubic downsampling operation. The light field SR models trained on the simulated pairs may exhibit poor performance when applied to real LR light fields where the degradations deviate from the simulated ones [61]. In this paper, we collect the first real-world paired LR-HR light field SR dataset using a Lytro ILLUM camera. We capture light fields at multiple focal lengths, providing a general benchmark for real-world light field SR.

3. Capturing the LytroZoom Dataset

Our goal is to collect a real-world light field SR dataset consisting of paired LR-HR light fields. This is challenging, as it requires accurately aligned LR-HR sequences of the same scene. To address this challenge, we use the Lytro

ILLUM camera, which can directly display the actual focal lengths during shooting. We capture city scenes printed on postcards and outdoor static objects as subjects.

We define a light field captured at 250 mm focal length as the HR ground-truth and the ones captured at 120mm and 60mm focal lengths as the $\times 2$ and $\times 4$ LR observations [5, 6]. We standardize our data collection process by shooting postcards at the lowest ISO setting and maintaining a constant white balance and aperture size. For outdoor scenes, we adjust the ISO value to ensure optimal exposure and minimize noise. The camera is fixed on a tripod for stabilization. Without post-processing procedures, such as color correction and histogram equalization, we decode the captured raw light field data using the light field toolbox [12, 13]. This results in $15 \times 15 \times 625 \times 434$ light field pairs. However, spatial misalignment, intensity variation, and color mismatching may exist in different views due to uncontrollable changes during lens zooming. Therefore, we use the alignment algorithm [5] view-by-view to rectify LR light fields iteratively to preserve view-dependent degradations. We then center-crop light field pairs to mitigate lens distortion and the vignetting effect. Please refer to Fig. 3 for the entire pipeline.

After careful shooting, decoding, rectifying, cropping, and selection, we collect a dataset with 94 city scenes printed on postcards and 63 outdoor static scenes. We provide pixel-wise aligned light field pairs (*i.e.*, ground-truth, $\times 2$ and $\times 4$ LR observations) with the resolution of $5 \times 5 \times 456 \times 320$ (LytroZoom-P) and $5 \times 5 \times 608 \times 416$ (LytroZoom-O). Fig. 2 shows samples from the dataset. We randomly partition LytroZoom-P into 63 scenes for training, 17 for validation, and the remaining 15 scenes for testing. Similarly, we use 55 scenes for training and the remaining 10 scenes for testing in LytroZoom-O.

4. Omni-Frequency Projection Network

In Sec. 3, we have collected LytroZoom, containing diverse contents captured by a Lytro ILLUM camera. Therefore we have access to the LR-HR pairs for training light

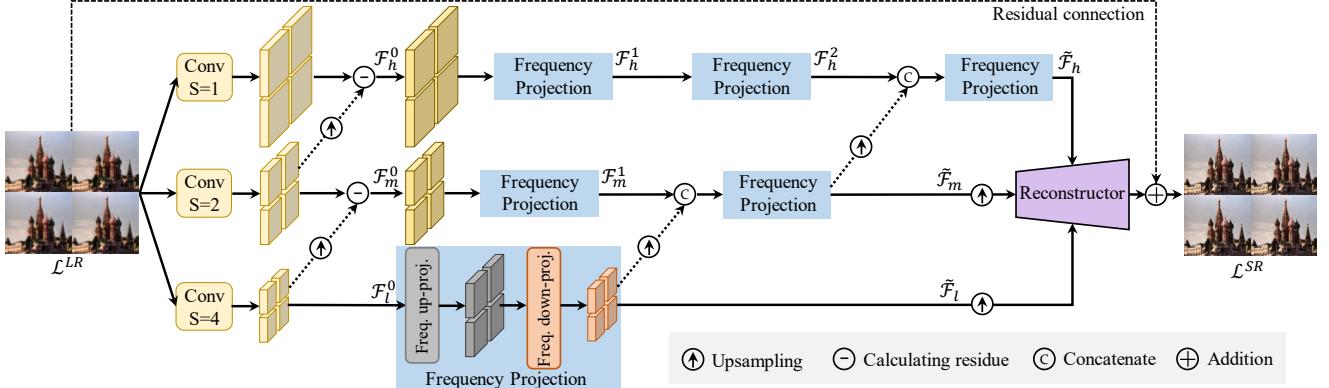


Figure 4. Architecture of our proposed OFPNet, which consists of three main components: (a) the omni-frequency decomposition which employs three convolutional layers to decompose the input real LR light field \mathcal{L}^{LR} in frequency domain, (b) the frequency projection to enhance texture details across different frequency components. (c) a reconstructor aims at aggregating the omni-frequency components for the HR light field reconstruction. The reconstructed \mathcal{L}^{SR} is obtained by adding the predicted residual to an LR light field.

field SR networks. However, existing light field SR methods do not entirely account for the characteristics of real-degraded light fields. To address this, we propose OFPNet for real-world light field SR, which can model real degradations in real-world light field SR.

4.1. Overview

Following [62, 64, 67, 72, 74, 78], we super-resolve the Y channel images, leaving Cb and Cr channel images being bicubic upscaled for light field SR. Without considering the channel dimension and given an LR light field $\mathcal{L}^{LR} \in \mathbb{R}^{U \times V \times H \times W}$ with few details and textures, we aim at generating an HR light field $\mathcal{L}^{SR} \in \mathbb{R}^{U \times V \times H \times W}$ with more details, which should be close to the ground-truth $\mathcal{L}^{GT} \in \mathbb{R}^{U \times V \times H \times W}$. U and V represent angular dimensions, and H and W represent spatial dimensions.

Inspired by recent progress in real-world single image SR, the degradation exists in all frequency components [15, 28, 36, 49, 85], we propose OFPNet, in which omni-frequency components are considered for real-world light field SR. In OFPNet, we first decompose the input LR light field \mathcal{L}^{LR} into high-frequency, middle-frequency, and low-frequency components, *i.e.*, \mathcal{F}_h , \mathcal{F}_m , and \mathcal{F}_l . Then the FP operations are utilized on three frequency branches to enhance frequency representations. These branches are interacted to enhance the cross-frequency representations. After we obtain the enhanced frequency features $\tilde{\mathcal{F}}_h$, $\tilde{\mathcal{F}}_m$, and $\tilde{\mathcal{F}}_l$, we feed them to the reconstructor to generate \mathcal{L}^{SR} .

4.2. Omni-Frequency Decomposition

To obtain the informative omni-frequency representation for real-world light field SR, we first decompose \mathcal{L}^{LR} into different frequency components. We utilize the learnable spatial downsampling operations to decompose frequency components in the feature domain, which is in a spirit similar to the octave convolution [1, 7].

As shown in Fig. 4, we first downsample \mathcal{L}^{LR} by a convolution layer with stride = 4 to get the corresponding low-frequency component \mathcal{F}_l . Then we obtain the middle frequency component \mathcal{F}_m by removing \mathcal{F}_l from the corresponding original feature, which is downsampled with a stride = 2 convolutional layer. Similarly, to get the high frequency component \mathcal{F}_h , we remove the downsampled feature with stride = 2 from the feature extracted with stride = 1 convolutional layer. The whole process can be denoted as

$$\begin{aligned}\mathcal{F}_l &= \text{conv}_2(\text{conv}_2(\mathcal{L}^{LR})), \\ \mathcal{F}_m &= \text{conv}_2(\mathcal{L}^{LR}) - [\text{conv}_2(\text{conv}_2(\mathcal{L}^{LR}))] \uparrow_2, \\ \mathcal{F}_h &= \text{conv}(\mathcal{L}^{LR}) - [\text{conv}_2(\mathcal{L}^{LR})] \uparrow_2,\end{aligned}\quad (1)$$

where $\text{conv}(\cdot)$ denotes the convolution layer without downsampling and $\text{conv}_2(\cdot)$ denotes the convolution layer with stride = 2. $[\cdot] \uparrow_r$ means the bilinear upsampling operation with the factor = r .

4.3. Frequency Projection

The extracted frequency components face the inevitable information loss problem caused by irreversible convolutional layers. Inspired by the back-projection operation that produces an HR feature map and iteratively refines it through multiple upsampling and downsampling layers to learn nonlinear relationships between LR and HR images [20–22, 24], we introduce the FP operation to enhance frequency feature representations, making up for the information lost.

As shown in Fig. 4, the FP operation consists of the frequency up-projection unit (FUPU) and the frequency down-projection unit (FDPU), in which nonlinear relationships between LR and HR frequency features can be exploited iteratively. We first project the extracted frequency feature \mathcal{F}^{n-1} to corresponding HR representation U^{n-1} based on a frequency scale-up block

$$U^{n-1} = \text{Up}(\mathcal{F}^{n-1}), \quad (2)$$

where $\text{Up}(\cdot)$ denotes the frequency scale-up block. It first fuses multi-view information progressively using the residual blocks, in which the inter-view correlations can be exploited, then upsamples the fused feature by bilinear interpolation followed by a 1×1 convolutional layer [42]. Please refer to the supplementary document for the detailed structure. m denotes the number of FUPU.

Then we project the HR representation back to LR one and obtain the corresponding residuals e^{n-1} between the back-projected representation and original LR input

$$e^{n-1} = \text{Down}(U^{n-1}) - \mathcal{F}^{n-1}, \quad (3)$$

where $\text{Down}(\cdot)$ denotes the frequency scale-down block. It first reduces the resolution of U^{n-1} to the original one via a 4×4 convolutional layer with stride=2, followed by fusing multi-view information progressively.

Finally, we back-project the residual to the HR representation and eliminate the corresponding super-resolved representation errors to obtain the final output of FUPU

$$U^n = \text{Up}(e^{n-1}) + U^{n-1}. \quad (4)$$

The procedure for FDPU is similar to FUPU. FDPU aims to obtain refined LR frequency representations by projecting the previously updated HR frequency representation. Please see the supplementary document for more details.

We can enhance the representations of different frequency components thanks to the FP operations. In practice, however, the high-frequency component is relatively tricky to enhance [15, 69]. We, therefore, propose to enhance such challenging frequency components in a coarse-to-fine manner. Specifically, we encourage the interaction between different frequency components and progressively utilize the enhanced lower frequency representations to help the enhancement of higher frequency components by concatenating them together. The final enhanced frequency features can be denoted as

$$\begin{aligned} \tilde{\mathcal{F}}_l &= \text{FP}_1(\mathcal{F}_h), \\ \tilde{\mathcal{F}}_m &= \text{FP}_2(\text{conv}([\text{FP}_1(\mathcal{F}_m), \tilde{\mathcal{F}}_l])), \\ \tilde{\mathcal{F}}_h &= \text{FP}_3(\text{conv}([\text{FP}_2(\text{FP}_1(\mathcal{F}_h)), \tilde{\mathcal{F}}_m])), \end{aligned} \quad (5)$$

where $\text{FP}_n(\cdot)$ denotes the n -th FP operation, and $\text{conv}(\cdot)$ here aims at reducing the channel dimensions.

4.4. Reconstructor

We feed the enhanced frequency features to the reconstructor to generate the super-resolved results. We first concatenate $\tilde{\mathcal{F}}_l, \tilde{\mathcal{F}}_m, \tilde{\mathcal{F}}_h$ along the channel dimension, followed by a convolutional layer to reduce the channel number. The concatenated feature is further fed to the Feature Blending Module (FBM) and the Upsampling Module [42]. Note that we remove the pixel shuffling operation because the LR-HR pairs have the same spatial resolution.

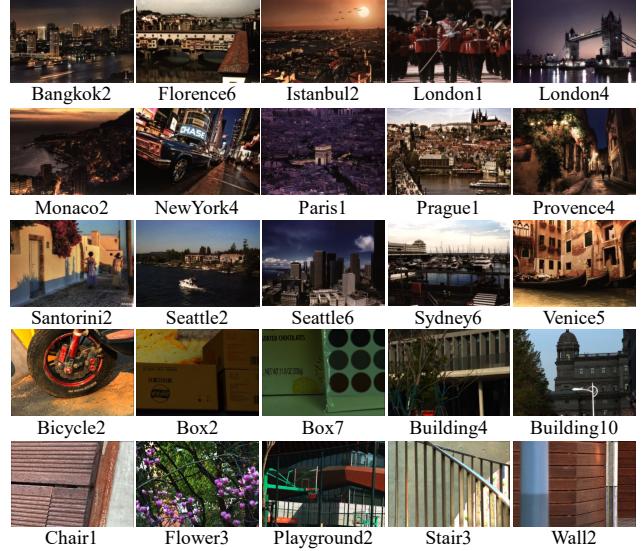


Figure 5. Thumbnails of 15 test scenes from LytroZoom-P (first three rows) and 10 test scenes from LytroZoom-O (last two rows). We show the ground-truth central view images here.

The L_1 -norm loss function is employed to minimize the pixel-wise distance between the generated HR light field \mathcal{L}^{SR} and the ground-truth \mathcal{L}^{GT} .

$$L(\mathcal{L}^{GT}, \mathcal{L}^{SR}) = \|\mathcal{L}^{GT} - \mathcal{L}^{SR}\|. \quad (6)$$

5. Experiments

5.1. Experimental Settings

Inference settings. PSNR and SSIM (the higher, the better) are adopted to evaluate the reconstruction accuracy. Following [62, 64, 67, 72, 74, 78], the light field SR results are evaluated using PSNR and SSIM indices on the Y channel in the YCbCr space. We evaluate the performance of different networks under the settings of $\times 2$ and $\times 4$ light field SR. To compare the angular consistency of the reconstructed HR results, the epipolar plane images (EPIS) are visualized for quantitative comparison in this paper.

Selected baseline methods. We select three representative and advanced light field SR networks, *i.e.*, InterNet [64], DPT [58], and IINet [42], as our main baselines. Note that, we find that ATO [29] and DistgSSR [63] cannot converge on LytroZoom, so we do not include them in the comparison. We follow the same experimental settings reported in their paper and retrain these networks based on their publicly available codes. Note that, we do not compare methods that require large memory consumptions during the training stage (*e.g.*, LFT [39]). We also exclude single image SR methods because previous work [62, 64, 67, 72, 74, 78] has proved that these methods cannot generate results with high fidelity and good angular consistency.

Implementation and training details of OFPNet. In our implementation, the channel number is set to 32 unless oth-

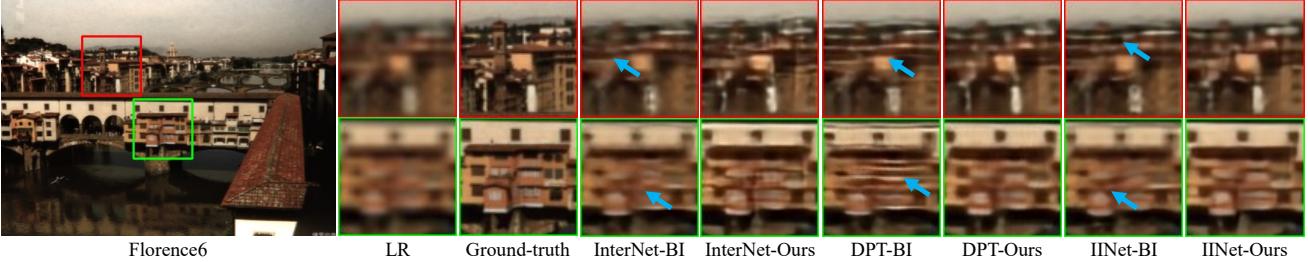


Figure 6. Visual comparisons ($\times 4$ SR) of different models (trained on the BI and LytroZoom-P datasets) on the LytroZoom-P testset.

Table 1. Average PSNR (dB) and SSIM results on the LytroZoom-P testset by different methods. “BI” indicates the method is trained on bicubic-downsampled datasets, and “LZ-P” indicates the method is trained on LytroZoom-P.

Metric	Scale	InterNet		DPT		IIINet		OFPNet
		BI	LZ-P	BI	LZ-P	BI	LZ-P	
PSNR	$\times 2$	31.55	38.78	31.55	38.65	31.88	38.78	38.89
	$\times 4$	26.12	29.60	25.63	29.43	26.04	29.82	30.11
SSIM	$\times 2$	0.9138	0.9764	0.9145	0.9754	0.9170	0.9771	0.9779
	$\times 4$	0.7685	0.8626	0.7569	0.8560	0.7621	0.8706	0.8786

erwise specified. We utilize the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Each mini-batch consists of 2 samples with 72×72 patches for $\times 2$ light field SR and 4 samples with 64×64 patches for $\times 4$ light field SR. We first train OFPNet on the LytroZoom-P dataset and then fine-tune the pretrained OFPNet on LytroZoom-O. The initial learning rate is set to $1e - 4$, and we reduce it by a factor of 0.5 every 2,000 epochs until 8,000 epochs during the training stage. During the fine-tuning stage, the initial learning rate is also set to $1e - 4$, and we reduce it by a factor of 0.5 every 1,000 epochs until 5,000 epochs. OFPNet is trained and fine-tuned on two NVIDIA GTX 1080Ti GPUs.

5.2. Simulated Datasets v.s. LytroZoom

To demonstrate the advantages of the LytroZoom dataset, we conduct experiments to compare the performance of light field SR models trained on simulated datasets and the LytroZoom-P dataset. We employ the mixed light field datasets [64] with the angular resolution of 5×5 to generate simulated $\times 2$ and $\times 4$ light field pairs with the bicubic degradation (BI). We train InterNet [64], IIINet [42], and DPT [58] on BI and LytroZoom-P for each of the two scaling factors ($\times 2$ and $\times 4$). Since the resolution of the LR-HR pair is the same in LytroZoom-P, we upsample the LR inputs of BI bicubically and feed the pre-upsampled inputs to the light field SR networks. We then attach the de-subpixel layer [56] at the beginning of each network to achieve an efficient inference.

We train three networks on BI and LytroZoom-P, respectively, and calculate the average PSNR and SSIM values on the LytroZoom-P testset. As shown in Table 1, models trained on our LytroZoom-P dataset obtain significantly better performance than those trained on the BI dataset for both scaling factors. Specifically, the results of the models trained on the BI dataset are even worse than the LR

Table 2. Average PSNR (dB) and SSIM results on LytroZoom-O testset by different methods (pretrained on LytroZoom-P and fine-tuned on LytroZoom-O).

Metric	Scale	InterNet	DPT	IIINet	OFPNet
PSNR	$\times 2$	30.15	30.08	30.28	30.79
	$\times 4$	27.88	27.95	28.22	28.91
SSIM	$\times 2$	0.8738	0.8603	0.8737	0.8863
	$\times 4$	0.7627	0.7620	0.7770	0.7991

observations. The reason is apparent: the networks trained with simulated LR-HR light field pairs inevitably get disastrous results when facing complex degradation in real-world scenes. In Fig. 6, we visualize the super-resolved central view images obtained by different models. As can be seen, results generated by models trained on the BI dataset tend to have blurring edges with obvious artifacts. On the contrary, models trained on LytroZoom-P generate clearer results.

5.3. Baseline Methods v.s. OFPNet

We compare our proposed OFPNet with three selected baseline methods. All models are trained/fine-tuned and tested on LytroZoom-P and LytroZoom-O.

LytroZoom-P. As is shown in Table 1, our OFPNet earns the highest PSNR and SSIM values. For example, one can see that OFPNet surpasses IIINet [42], the state-of-the-art light field SR method, by 0.29dB/0.0080 on $\times 4$ SR in terms of PSNR/SSIM. Fig. 7 shows the super-resolved central view images for qualitative comparison. In terms of visual quality, OFPNet beats previous methods, providing fine details without introducing unappealing artifacts in general. For example, as seen in Fig. 7, baseline methods can hardly restore the external shape of the buildings in Bangkok2. In contrast, our proposed OFPNet can generate results with vivid details and patterns.

LytroZoom-O. As is shown in Table 2, on the LytroZoom-O testset, our OFPNet also earns the highest PSNR and SSIM values. Fig. 8 shows the super-resolved central view images from LytroZoom-O for qualitative comparison. OFPNet can generate results with fine details.

5.4. Generalization Tests

Our LytroZoom-trained light field SR models exhibit robust generalization capabilities, both in terms of content and device. Specifically, our LytroZoom-P-trained models perform well on a scene captured at 200 mm focal length,

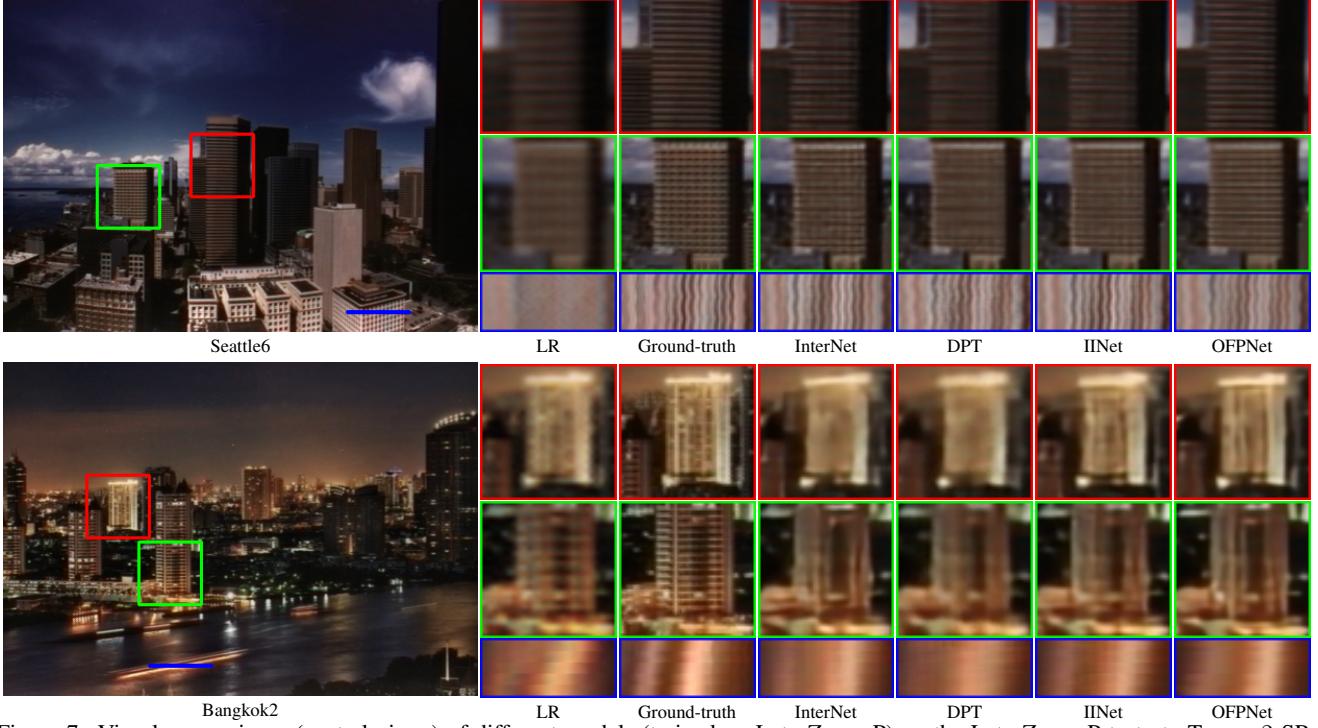


Figure 7. Visual comparisons (central views) of different models (trained on LytroZoom-P) on the LytroZoom-P testset. Top: $\times 2$ SR. Bottom: $\times 4$ SR. Please zoom in for better visualization and best viewed on the screen.

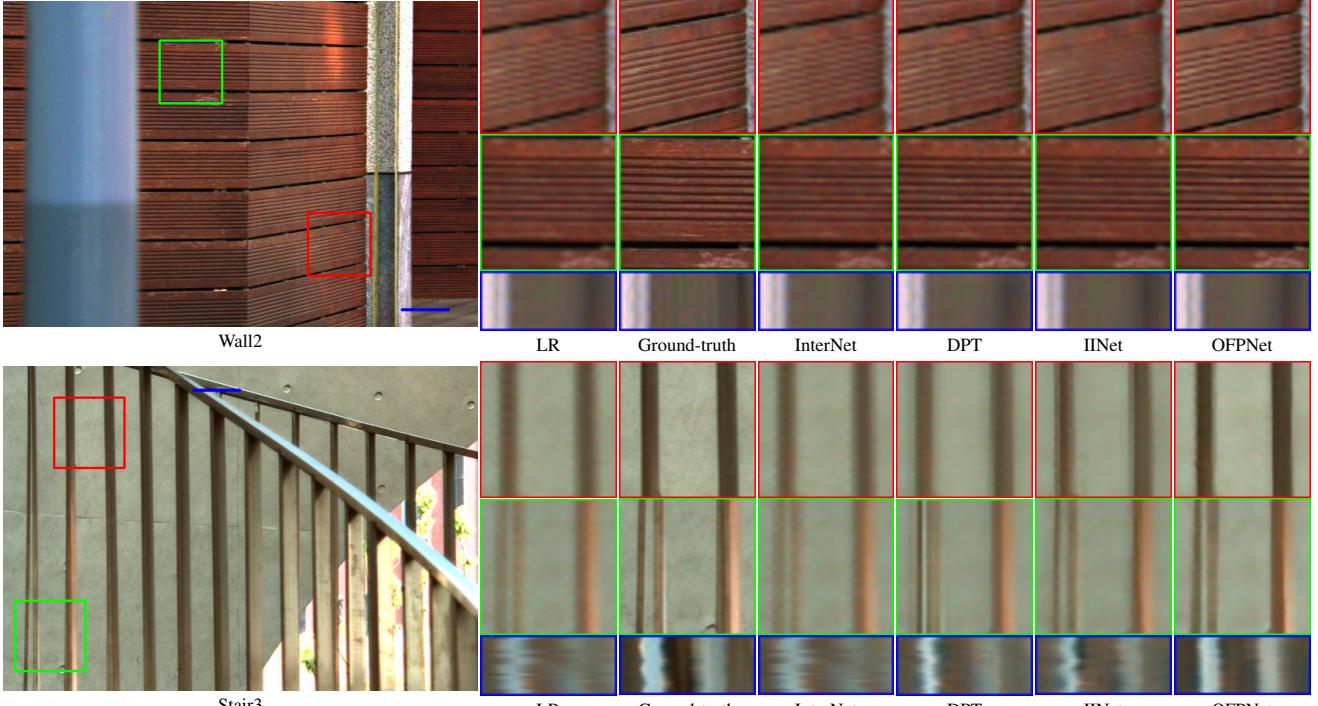


Figure 8. Visual comparisons (central views) of different models (fine-tuned on LytroZoom-O) on the LytroZoom-O testset. Top: $\times 2$ SR. Bottom: $\times 4$ SR. Please zoom in for better visualization and best viewed on the screen.

as demonstrated in Fig. 9, despite being trained on indoor scenes printed on postcards. For the device generalization, as shown in Fig. 10, the light field SR models trained on

the dataset captured by a Lytro ILLUM camera can be readily applied to different devices such as Gantry (we super-resolve the input light field directly from the scene *Lego*-



Figure 9. Visual comparisons (central views) of different models on real-world light field scenes ($\times 2$ SR). LZ-P is short for LytroZoom-P.

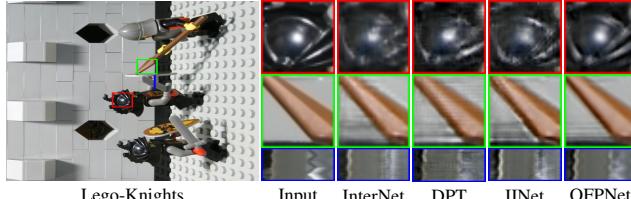


Figure 10. Visual comparisons (central views) of different models on real-world light field scene captured by Gantry.

Table 3. Analysis of the OFPNet on $\times 4$ SR on LytroZoom-P.

Frequency decomposition			Frequency Projection		PSNR
\mathcal{F}_l	\mathcal{F}_m	\mathcal{F}_h	Interactions	FP operation	
\times	\times	\checkmark	-	-	29.87
\times	\checkmark	\checkmark	-	-	29.98
\checkmark	\checkmark	\checkmark	-	-	30.11
-	-	-	\times	\times	29.60
-	-	-	\checkmark	\times	29.86
-	-	-	\times	\checkmark	29.99
-	-	-	\checkmark	\checkmark	30.11

Knights in STFgantry [55]).

5.5. Model Analysis

Investigation of the frequency decomposition. We investigate different extracted frequency components in OFPNet. We have added several residual blocks after the extracted frequency features while removing the corresponding components in Table 3 to ensure the parameters remain unchanged. As shown in Table 3, our results demonstrate that incorporating additional frequency components in OFPNet leads to improved performance in terms of PSNR, with gains of 0.24 dB and 0.13 dB when considering higher frequency components. These results suggest that the omni-frequency components play a crucial role in real-world light field SR. Further details and analysis can be found in the supplementary document.

Investigation of frequency projection. When we simultaneously remove the interactions between frequency components and the FP operations (we replace the FP operations with residual blocks of the same parameters), we only get a result of 29.60dB in terms of PSNR. When we add interaction and FP operations, the results improve by 0.26dB and 0.39dB, respectively, indicating the importance of these two designs. The best result can be obtained using these two designs simultaneously (30.11dB) on LytroZoom-P.

6. Discussion

Despite the encouraging performance as shown above, the LytroZoom dataset still has certain limitations. (1) The use of a single data collection device, specifically a Lytro ILLUM camera, limits the generalizability of models trained on LytroZoom to other cameras. While these models can perform well on the Gantry dataset, there may still be a domain shift when applied to light fields captured by cameras with different baselines, resulting in artifacts (as seen in Fig. 10). To address this limitation, future work can expand the LytroZoom dataset to include light fields captured by other types of cameras, such as camera arrays and Gantry. (2) Minor distortions that cannot be rectified. The registration step [5] could alleviate the distortions caused by different FoVs, yet minor misalignment and luminance/color differences exist between the LR-HR light fields. It is likely due to these minor distortions that cause the non-convergence of ATO [29] and DistgSSR [63] during the training stage [71]. We will investigate new training strategies. (3) Inadequate benchmark experiments. We will extend LytroZoom with more scaling factors and conduct experiments on scale-arbitrary real-world light field SR.

7. Conclusion

In this paper, we have collected the first real-world paired LR-HR light field SR dataset, *i.e.*, LytroZoom, with authentic degradation for real-world light field SR. Specifically, a Lytro ILLUM camera is used to collect accurate pixel-wise aligned LR and HR light field pairs on 94 city scenes printed on postcards (LytroZoom-P) and 63 outdoor static objects (LytroZoom-O). A novel OFPNet is proposed to efficiently solve the real-world light field SR problem by considering omni-frequency information and enhancing cross-frequency representations. Our extensive experiments validate that the models trained on our LytroZoom dataset can lead to much better real-world light field SR results than trained on simulated datasets. They are also readily generalizable to diverse content and devices. We believe that LytroZoom presents a valuable opportunity for the research community to further explore real-world light field SR.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

- [1] Mohammad Akbari, Jie Liang, Jingning Han, and Chengjie Tu. Generalized octave convolutions for learned multi-frequency image compression. *arXiv preprint arXiv:2002.10032*, 2020. 4
- [2] Martin Alain and Aljosa Smolic. Light field super-resolution via lfbm5d sparse coding. In *ICIP*, 2018. 1, 2
- [3] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *NeurIPS*, 32, 2019. 2
- [4] James R Bergen and Edward H Adelson. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1:8, 1991. 1
- [5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 2, 3, 8
- [6] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, 2019. 2, 3
- [7] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *ICCV*, 2019. 4
- [8] Zhen Cheng, Zhiwei Xiong, Chang Chen, and Dong Liu. Light field super-resolution: A benchmark. In *CVPRW*, 2019. 2
- [9] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *CVPR*, 2021. 2, 3
- [10] Zhen Cheng, Zhiwei Xiong, and Dong Liu. Light field super-resolution by jointly exploiting internal and external similarities. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2604–2616, 2019. 1, 3
- [11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 2
- [12] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *CVPR*, 2013. 3
- [13] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams. Linear volumetric focus for light field cameras. *ACM Transactions on Graphics (TOG)*, 34(2), Feb. 2015. 3
- [14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2, 3
- [15] Manuel Fritzsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCVW*, 2019. 2, 4, 5
- [16] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *ICCV*, 2021. 2
- [17] Guangwei Gao, Wenjie Li, Juncheng Li, Fei Wu, Huimin Lu, and Yi Yu. Feature distillation interaction weighting network for lightweight image super-resolution. In *AAAI*, 2022. 2
- [18] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, 2019. 2
- [19] M Shahzeb Khan Gul and Bahadir K Gunturk. Spatial and angular resolution enhancement of light fields using convolutional neural networks. *IEEE Transactions on Image Processing*, 27(5):2146–2159, 2018. 1
- [20] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 4
- [21] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019. 4
- [22] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projectinetworks for single image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4323–4337, 2020. 4
- [23] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*, 2016. 3
- [24] Mengshun Hu, Kui Jiang, Liang Liao, Jing Xiao, Junjun Jiang, and Zheng Wang. Spatial-temporal space hand-in-hand: Spatial-temporal video super-resolution via cycle-projected mutual learning. In *CVPR*, 2022. 4
- [25] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *NeurIPS*, 2020. 2
- [26] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, 2019. 2
- [27] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, 2020. 2
- [28] Xiaozhong Ji, Guangpin Tao, Yun Cao, Ying Tai, Tong Lu, Chengjie Wang, Jilin Li, and Feiyue Huang. Frequency consistent adaptation for real world super resolution. In *AAAI*, 2021. 4
- [29] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *CVPR*, 2020. 1, 3, 5, 8
- [30] Neel Joshi, Shai Avidan, Wojciech Matusik, and David J Kriegman. Synthetic aperture tracking: tracking through occlusions. In *ICCV*, 2007. 1
- [31] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 2
- [32] Soo Ye Kim, Hyeonjun Sim, and Munchurl Kim. Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *CVPR*, 2021. 2
- [33] 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. 3
- [34] Juncheng Li, Faming Fang, Jiaqian Li, Kangfu Mei, and Guiyu Zhang. Mdcn: Multi-scale dense cross network for image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2547–2561, 2020. 2

- [35] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, 2018. 2
- [36] Xin Li, Xin Jin, Tao Yu, Simeng Sun, Yingxue Pang, Zhizheng Zhang, and Zhibo Chen. Learning omni-frequency region-adaptive representations for real image super-resolution. In *AAAI*, 2021. 4
- [37] Chia-Kai Liang and Ravi Ramamoorthi. A light transport framework for lenslet light field cameras. *ACM Transactions on Graphics (TOG)*, 34(2):1–19, 2015. 2
- [38] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 2
- [39] 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
- [40] 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
- [41] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2
- [42] 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. 1, 5, 6
- [43] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 2
- [44] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *ICCVW*, 2019. 2
- [45] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, 2021. 2
- [46] 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. 1, 3
- [47] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford University, 2005. 1
- [48] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, 2020. 2
- [49] Yingxue Pang, Xin Li, Xin Jin, Yaojun Wu, Jianzhao Liu, Sen Liu, and Zhibo Chen. Fan: Frequency aggregation network for real image super-resolution. In *ECCV*, 2020. 2, 4
- [50] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. Stanford lytro light field archive, 2016. 3
- [51] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 3
- [52] Mattia Rossi and Pascal Frossard. Graph-based light field super-resolution. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, 2017. 1, 2
- [53] Hao Sheng, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui. Urbanlf: a comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7880–7893, 2022. 3
- [54] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017. 2
- [55] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7), 2008. 3, 8
- [56] Thang Vu, Cao Van Nguyen, Trung X Pham, Tung M Luu, and Chang D Yoo. Fast and efficient image quality enhancement via desubpixel convolutional neural networks. In *EC-CVW*, 2018. 6
- [57] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, 2021. 2
- [58] Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Detail preserving transformer for light field image super-resolution. In *AAAI*, 2022. 1, 3, 5, 6
- [59] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *ICCV*, 2021. 2
- [60] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *ICCV*, 2015. 1
- [61] Yingqian Wang, Zhengyu Liang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Learning a degradation-adaptive network for light field image super-resolution. *arXiv preprint arXiv:2206.06214*, 2022. 3
- [62] 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. 1, 3, 4, 5
- [63] 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. 1, 2, 3, 5, 8
- [64] 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. 1, 3, 4, 5, 6
- [65] Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, and Yulan Guo. Deocnet: Learning to see through foreground occlusions in light fields. In *WACV*, 2020. 1
- [66] 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

- [67] 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. 1, 4, 5
- [68] 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. Citeseer, 2013. 3
- [69] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 2020. 2, 5
- [70] Zeyu Xiao, Yutong Liu, Ruisheng Gao, and Zhiwei Xiong. Cutmib: Boosting light field super-resolution via multi-view image blending. In *CVPR*, 2023. 1
- [71] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *ICCV*, 2021. 8
- [72] 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. 1, 3, 4, 5
- [73] 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, 3
- [74] Yan Yuan, Ziqi Cao, and Lijuan Su. Light-field image super-resolution using a combined deep cnn based on epi. *IEEE Signal Processing Letters*, 25(9):1359–1363, 2018. 1, 3, 4, 5
- [75] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, 2018. 2
- [76] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 2
- [77] 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. 1
- [78] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *CVPR*, 2019. 1, 3, 4, 5
- [79] Shuo Zhang, Zeqi Shen, and Youfang Lin. Removing foreground occlusions in light field using micro-lens dynamic filter. In *IJCAI*, 2021. 1
- [80] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 1
- [81] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *CVPR*, 2019. 2
- [82] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2
- [83] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 2
- [84] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *ICCV*, 2019. 2
- [85] Yuanbo Zhou, Wei Deng, Tong Tong, and Qinquan Gao. Guided frequency separation network for real-world super-resolution. In *CVPRW*, 2020. 4