Light Field Spatial Super-resolution via Deep Combinatorial Geometry Embedding and Structural Consistency Regularization

Jing Jin[1]*, Junhui Hou[1] †, Jie Chen[2] , Sam Kwong[1]
[1]City University of Hong Kong [2] Hong Kong Baptist University
jingjin25-c@my.cityu.edu.hk, {jh.hou,cssamk}@cityu.edu.hk, chenjie@comp.hkbu.edu.hk

Abstract

Light field (LF) images acquired by hand-held devices usually suffer from low spatial resolution as the limited sampling resources have to be shared with the angular dimension. LF spatial super-resolution (SR) thus becomes an indispensable part of the LF camera processing pipeline. The high-dimensionality characteristic and complex geometrical structure of LF images make the problem more challenging than traditional single-image SR. The performance of existing methods is still limited as they fail to thoroughly explore the coherence among LF views and are insufficient in accurately preserving the parallax structure of the scene. In this paper, we propose a novel learning-based LF spatial SR framework, in which each view of an LF image is first individually super-resolved by exploring the complementary information among views with combinatorial geometry embedding. For accurate preservation of the parallax structure among the reconstructed views, a regularization network trained over a structure-aware loss function is subsequently appended to enforce correct parallax relationships over the intermediate estimation. Our proposed approach is evaluated over datasets with a large number of testing images including both synthetic and real-world scenes. Experimental results demonstrate the advantage of our approach over state-of-the-art methods, i.e., our method not only improves the average PSNR by more than 1.0 dB but also preserves more accurate parallax details, at a lower computational cost.

1. Introduction

4D light field (LF) images differ from conventional 2D images as they record not only intensities but also directions of light rays [38]. The rich information enables a wide range of applications, such as 3D reconstruction [15, 39, 27, 40], refocusing [9], and virtual reality [13, 36]. LF images can be conveniently captured with commercial micro-lens based cameras [1, 2] by encoding the 4D LF into a 2D photo detector. However, due to the limited resolution of the sensor, recorded LF images always suffer from low spatial resolution. Therefore, LF spatial super-resolution (SR) is highly necessary for further applications.

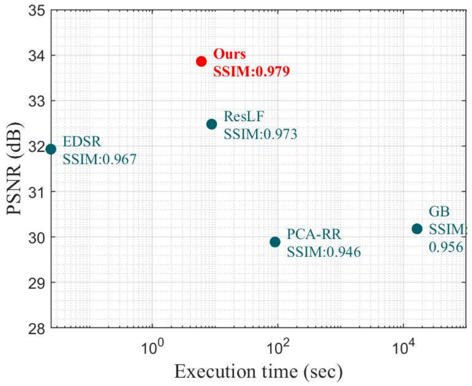


Figure 1. Comparisons of the running time (in second) and reconstruction quality (PSNR/SSIM) of different methods. The running time is the time for super-resolving an LF image of spatial resolution 94 × 135 and angular resolution 7 × 7 with the scale factor equal to 4. The PSNR/SSIM value refers to the average over 57 LF images in Stanford Lytro Archive (Occlusions) dataset.

Some traditional methods for LF spatial SR have been proposed [31, 23, 20]. Due to the high dimensionality of LF data, the reconstruction quality of these methods is quite limited. Recently, some learning-based methods [34, 28, 37] have been proposed to address the problem of 4D LF spatial SR via data-driven training. Although these methods have improved both performance and efficiency, there are two problems unsolved yet. That is, the complementary information within all views is not fully utilized, and the structural consistency of the reconstruction is not well preserved (see more analyses in Sec. 3).

In this paper, we propose a learning-based method for

*This work was supported in part by the Hong Kong Research Grants Council under grants 9048123 and 9042820, and in part by the Basic Research General Program of Shenzhen Municipality under grants JCYJ20190808183003968.

†Corresponding author.

LF spatial SR, focusing on addressing the two problems of complete complementary information fusion and LF parallax structure preservation. As shown in Fig. 3, our approache consists of two modules, i.e., an All-to-One SR via combinatorial geometry embedding and a structural consistency regularization module. Specifically, the All-to-One SR module separately super-resolves individual views by learning combinatorial correlations and fusing the complementary information of all views, giving an intermediate super-resolved LF image. The regularization module exploits the spatial-angular geometry coherence among the intermediate result, and enforces the structural consistency in the high-resolution space. Extensive experimental results on both real-world and synthetic datasets demonstrate the advantage of our proposed method. That is, as shown in Fig. 1, our method produces much higher PSNR/SSIM at a higher speed, compared with state-of-the-art methods.

2. Related Work

Two-plane representation of 4D LFs. The 4D LF is commonly represented using two-plane parameterization. Each light ray is determined by its intersections with two parallel planes, i.e., a spatial plane (x, y) and a angular plane (u, v). Let $L(\mathbf{x}, \mathbf{u})$ denote a 4D LF image, where $\mathbf{x} = (x, y)$ and $\mathbf{u} = (u, v)$. A view, denoted as $L_{\mathbf{u}^*} = L(\mathbf{x}, \mathbf{u}^*)$, is a 2D slice of the LF image at a fixed angular position $\mathbf{u}^*$. The views with different angular positions capture the 3D scene from slightly different viewpoints.

Under the assumption of Lambertian, projections of the same scene point will have the same intensity at different views. This geometry relation leads to a particular *LF parallax structure*, which can be formulated as:

$$L_{\mathbf{u}}(\mathbf{x}) = L_{\mathbf{u}'}(\mathbf{x} + d(\mathbf{u}' - \mathbf{u})), \tag{1}$$

where d is the disparity of the point $L(\mathbf{x}, \mathbf{u})$. The most straightforward representation of the LF parallax structure is epipolar-plane images (EPIs). Specifically, each EPI is the 2D slice of the 4D LF at one fixed spatial and angular position, and consists of straight lines with different slops corresponding to scene points at different depth.

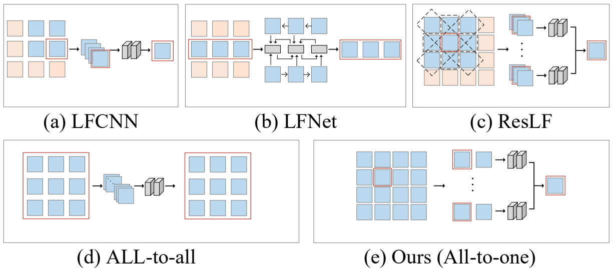
LF spatial SR. For single image, the inverse problem of SR is always addressed using different image statistics as priors [22]. As multiple views are available in LF images, the correlations between them can be used to directly constrain the inverse problem, and the complementary information between them can greatly improve the performance of SR. Existing methods for LF spatial SR can be classed into two categories: optimization-based and learning-based methods.

Traditional LF spatial SR methods physically model the relations between views based on estimated disparities, and then formulate SR as an optimization problem. Bishop and Favaro [4] first estimated the disparity from the LF image, and then used it to build an image formation model, which is employed to formulate a variantional Bayesian framework for SR. Wanner and Goldluecke [30, 31] applied structure tensor on EPIs to estimate disparity maps, which were employed in a variational framework for spatial and angular SR. Mitra and Veeraraghavan [20] proposed a common framework for LF processing, which models the LF patches using a Gaussian mixture model conditioned on their disparity values. To avoid the requirement of precise disparity estimation, Rossi and Frossard [23] proposed to regularize the problem using a graph-based prior, which explicitly enforces the LF geometric structure.

Learning-based methods exploit the cross-view redundancies and utilize the complementary information between views to learn the mapping from low-resolution to high-resolution views. Farrugia [8] constructed a dictionary of examples by 3D patch-volumes extracted from pairs of low-resolution and high-resolution LFs. Then a linear mapping function is learned using Multivariate Ridge Regression between the subspace of these patch-volumes, which is directly applied to super-resolve the low-resolution LF images. Recent success of CNNs in single image super-resolution (SISR) [6, 18, 29] inspired many learning-based methods for LF spatial SR. Yoon *et al*. [35, 34] first proposed to use CNNs to process LF data. They used a network with similar architecture of that in [6] to improve the spatial resolution of neighboring views, which were used to interpolate novel views for angular SR next. Wang *et al*. [28] used a bidirectional recurrent CNN to sequentially model correlations between horizontally or vertically adjacent views. The predictions of horizontal and vertical subnetworks are combined using the stacked generalization technique. Zhang *et al*. [37] proposed a residual network to super-resolve the view of LF images. Similar to [26], views along four directions are first stacked and fed into different branches to extract sub-pixel correlations. Then the residual information from different branches is integrated for final reconstruction. However, the performance of side views will be significantly degraded compared with the central view as only few views can be utilized, which will result in undesired inconsistency in the reconstructed LF images. Additionally, this method requires various models suitable for views at different angular positions, *e.g*., 6 models for a 7×7 LF image, which makes the practical storage and application harder. Yeung *et al*. [32] used the alternate spatial-angular convolution to super-resolve all views of the LF at a single forward pass.

3. Motivation

Given a low-resolution LF image, denoted as $L^{lr} \in \mathbb{R}^{H \times W \times M \times N}$, LF spatial SR aims at reconstructing a super-resolved LF image, close to the ground-truth high-



(a) LFCNN (b) LFNet (c) ResLF

(d) ALL-to-all (e) Ours (All-to-one)

Figure 2. Illustration of different network architectures for the fusion of view complementary information. (a) LFCNN [34], (b) LFNet [28], (c) ResLF [37], (d) an intuitive *All-to-All* fusion (see Sec. 3), and (e) our proposed *All-to-One* fusion. Colored boxes represent images or feature maps of different views. Among them, red-framed boxes are views to be super-resolved, and blue boxes are views whose information is utilized.

resolution LF image $L^{hr} \in \mathbb{R}^{\alpha H \times \alpha W \times M \times N}$, where $H \times W$ is the spatial resolution, $M \times N$ is the angular resolution, and α is the upsampling factor. We believe the following two issues are paramount for high-quality LF spatial SR: (1) thorough exploration of the complementary information among views; and (2) strict regularization of the *view-level* LF structural parallax. In what follows, we will discuss more about these issues, which will shed light on the proposed method.

(1) **Complementary information among views**. An LF image contains multiple observations of the same scene from slightly varying angles. Due to occlusion, non-Lambertian reflections, and other factors, the visual information is asymmetric among these observations. In other words, the information absent in one view may be captured by another one, hence all views are potentially helpful for high-quality SR.

Traditional optimization-based methods [30, 31, 23, 20] typically model the relationships among views using explicit disparity maps, which is expensive to compute. Moreover, inaccurate disparity estimation in occluded or non-Lambertian regions will induce artifacts and the correction of such artifacts is beyond the capabilities of these optimization-based models. Instead, recent learning-based methods, such as LFCNN [35], LFNet [28] and ResLF [37], explore the complementary information among views through data-driven training. Although these methods improve both the reconstruction quality and computational efficiency, the complementary information among views has not been fully exploited due to the limitation of their view fusion mechanisms. Fig. 2 shows the architectures of different view fusion approaches. LFCNN only uses neighbouring views in a pair or square, while LFNet only takes views in a horizontal and vertical 3D LF. ResLF considers 4D structures by constructing directional stacks, which leaves views not located at the "star" shape un-utilized.

Remark. An intuitive way to fully take advantage of the cross-view information is by stacking the images or features of all views, feeding them into a deep network, and predicting the high-frequency details for all views simultaneously. We refer to this method *All-to-All* in this paper. As illustrated in Fig. 2(d), this is a naive extension of the classical SISR networks [16]. However, this method will compromise unique details that only belong to individual views since it is the average error over all views which is optimized during network training. See the quantitative verification in Sec. 5.2. To the end, we propose a novel fusion strategy for LF SR, called *All-to-One* SR via *combinatorial geometry embedding*, which super-resolves each individual view by combining the information from all views.

(2) **LF parallax structure**. As the most important property of an LF image, the parallax structure should be well preserved after SR. Generally, existing methods promote the fidelity of such a structure by enforcing corresponding pixels to share similar intensity values. Specifically, traditional methods employ particular regularization in the optimization formulation, such as the low-rank [11] and graph-based [23] regularizer. Farrugia and Guillemot [7] first used optical flow to align all views and then super-resolve them simultaneously via an efficient CNN. However, the disparity between views need to be recovered by warping and inpainting afterwards, which will cause inevitable high-frequency loss. For most learning-based methods [28, 37], the cross-view correlations are only exploited in the low-resolution space, while the consistency in the high-resolution space is not well modeled. See the quantitative verification in Sec. 5.1.

Remark. We address the challenge of LF parallax structure preservation with a subsequent regularization module on the intermediate high-resolution results. Specifically, an additional network is applied to explore the spatial-angular geometry coherence in the high-resolution space, which models the parallax structure implicitly. Moreover, we use a structure-aware loss function defined on EPIs, which enforces not only view consistency but also models inconsistency on non-Lambertian regions.

4. The Proposed Method

As illustrated in Fig. 3, our approach consists of an All-to-One SR module, which super-resolves each view of an LF image individually by fusing the combinatorial embedding from all other views, and followed by a structural consistency regularization module, which enforces the LF parallax structure in the reconstructed LF image.

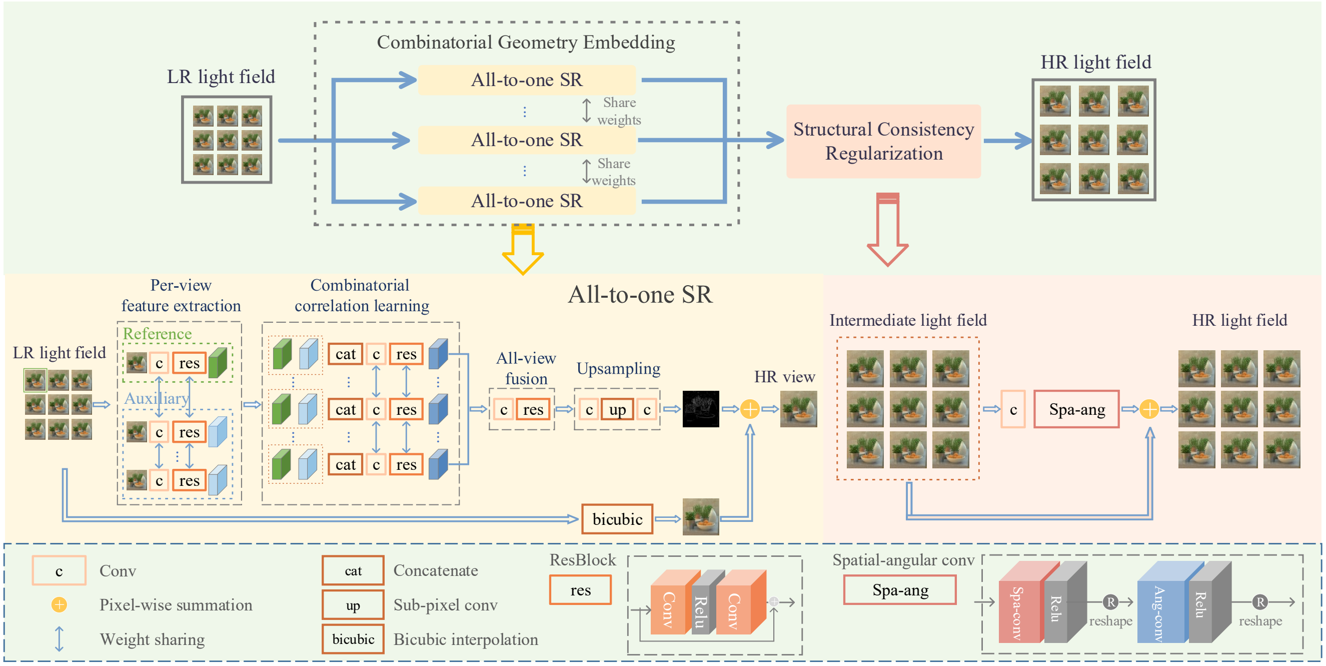


Figure 3. The flowchart of our proposed approach and illustration of the detailed architecture of the *All-to-One* SR and structural consistency regularization modules. The All-to-One SR module takes full advantage of the complementary information of all views of an LF image by learning their combinatorial correlations with the reference view. At the same time, the unique details of each individual view are also well retained. The structural consistency regularization module recovers the view consistency among the resulting intermediate LF image by exploring the spatial-angular relationships and a structure-aware loss.

4.1. All-to-One SR via Combinatorial Geometry Embedding

Let $L^{lr}_{\mathbf{u}_r}$ denote the reference view to be super-resolved. The remaining views of an LF image except $L^{lr}_{\mathbf{u}_r}$ are denoted as auxiliary views $\{L^{lr}_{\mathbf{u}_a}\}$. The All-to-One SR module focuses on extracting the complementary information from auxiliary views to assist the SR of the reference view. As shown in Fig. 3, there are four sub-phases involved, i.e., per-view feature extraction, combinatorial correlation learning, all-view fusion and upsampling.

Per-view feature extraction. We first extract deep features, denoted as $F^1_{\mathbf{u}}$, from all views separately, i.e.,

$$F^1_{\mathbf{u}} = f_1(L^{lr}_{\mathbf{u}}). \tag{2}$$

Inspired by the excellent performance of residual blocks [10, 16], which learn residual mappings by incoporating the self-indentity, we use them for deep feature extraction. The feature extraction process $f_1(\cdot)$ contains a convolutional layer followed by rectified linear units (ReLU), and n_1 residual blocks. The parameters of $f_1(\cdot)$ are shared across all views.

Combinatorial correlation learning. The geometric correlations between the reference view and auxiliary views vary with their angular positions $\mathbf{u}_r$ and $\mathbf{u}_a$. To enable our model to be compatible for all views with different $\mathbf{u}_r$ in the LF, we use the network $f_2(\cdot)$ to learn the correlations between the features of a pair of views $\{F^1_{\mathbf{u}_1}, F^1_{\mathbf{u}_2}\}$, where the angular positions $\mathbf{u}_1$ and $\mathbf{u}_2$ can be arbitrarily selected. Based on the correlations between $F^1_{\mathbf{u}_1}$ and $F^1_{\mathbf{u}_2}$, $f_2(\cdot)$ is designed to extract information from $F^1_{\mathbf{u}_2}$ and embed it into the features of $F^1_{\mathbf{u}_1}$. Here, $\mathbf{u}_1$ is set to be the angular position of the reference view, and $\mathbf{u}_2$ can be the position of any auxiliary view. Thus the output can be written as:

$$F^{2,\mathbf{u}_a}_{\mathbf{u}_r} = f_2(F^1_{\mathbf{u}_r}, F^1_{\mathbf{u}_a}), \tag{3}$$

where $F^{2,\mathbf{u}_a}_{\mathbf{u}_r}$ is the features of the reference view $L^{lr}_{\mathbf{u}_r}$ incorporated with the information of an auxiliary view $L^{lr}_{\mathbf{u}_a}$.

The network $f_2(\cdot)$ consists of a concatenation operator to combine the features $F^1_{\mathbf{u}_r}$ and $F^1_{\mathbf{u}_a}$ as inputs, and a convolutional layer followed by n_2 residual blocks. $f_2(\cdot)$'s ability of handling arbitrary pair of views is naturally learned by accepting the reference view and all auxiliary views in each training iteration.

All-view fusion. The output of $f_2(\cdot)$ is a stack of features with embedded geometry information from all auxiliary views. These features have been trained to align to the reference view, hence they can be fused directly. The fusion process can be formulated as:

$$F^3_{\mathbf{u}_r} = f_3(F^{2,\mathbf{u}_{a_1}}_{\mathbf{u}_r}, \cdots, F^{2,\mathbf{u}_{a_m}}_{\mathbf{u}_r}), \tag{4}$$

where $m = MN - 1$ is the number of auxiliary views.

Instead of concatenating all features together, we first combine them channel-wise, i.e. combine the feature maps at the same channel across all views. Then, all channel maps are used to extract deeper features. The network $f_3(\cdot)$ consists of one convolutional layer, n_3 residual blocks for channel-wise view fusion and n_4 residual blocks for channel fusion.

Upsampling. We use a similar architecture with residual learning in SISR [16]. To reduce the memory consumption and computational complexity, all feature learning and fusion are conducted in low-resolution space. The fused features are upsampled using the efficient sub-pixel convolutional layer [25], and a residual map is then reconstructed by a subsequent convolutional layer $f_4(\cdot)$. The final reconstruction is produced by adding the residual map with the upsampled image:

$$L^{sr}_{\mathbf{u}_r} = f_4(U_1(F^3_{\mathbf{u}_r})) + U_2(L^{lr}_{\mathbf{u}_r}), \tag{5}$$

where $U_1(\cdot)$ is the sub-pixel convolutional layer and $U_2(\cdot)$ is the bicubic interpolation process.

Loss function. The objective of the All-to-One SR module is to super-resolve the reference view individually $\widehat{L}^{sr}_{\mathbf{u}_r}$ to approach the ground truth high-resolution image $L^{hr}_{\mathbf{u}_r}$. We use the ℓ_1 error between them to define the loss function:

$$\ell_v = ||\widehat{L}^{sr}_{\mathbf{u}_r} - L^{hr}_{\mathbf{u}_r}||_1. \tag{6}$$

4.2. Structural Consistency Regularization

We apply structural consistency regularization on the intermediate results by the All-to-One SR module. This regularization module employs the efficient alternate spatial-angular convolution to implicitly model cross-view correlations among the intermediate LF images. In addition, a structure-aware loss function defined on EPIs is used to enforce the structural consistency of the final reconstruction.

Efficient alternate spatial-angular convolution. To regularize the LF parallax structure, an intuitive method is using the 4D or 3D convolution. However, 4D or 3D CNNs will result in significant increase of the parameter number and computational complexity. To improve the efficiency, but still explore the spatial-angular correlations, we adopt the alternate spatial-angular convolution [21, 32, 33], which handles the spatial and angular dimensions in an alternating manner with the 2D convolution.

In our regularization network, we use n_5 layers of alternate spatial-angular convolutions. Specifically, for the intermediate results $\widehat{L}^{sr} \in \mathbb{R}^{\alpha H \times \alpha W \times M \times N}$, we first extract features from each view separately and construct a stack of spatial views, i.e., $F_s \in \mathbb{R}^{\alpha H \times \alpha W \times c \times MN}$, where c is the number of feature maps. Then we apply 2D spatial convolutions on F_s. The output features are reshaped to the stacks of angular patches, i.e., $F_a \in \mathbb{R}^{M \times N \times c \times \alpha^2 HW}$, and then angular convolutions are applied. Afterwards, the features are reshaped for spatial convolutions, and the previous 'Spatial Conv-Reshape-Angular Conv-Reshape' process repeats n_5 times.

Table 1. The datasets used for evaluation.

	Dataset	category	#scenes
Real-world	Stanford Lytro Archive [3]	General Occlusions	57 51
	Kalantari *et al.* [14]	testing	30
Synthetic	HCI new [12] Inria Synthetic [24]	testing DLFD	4 39

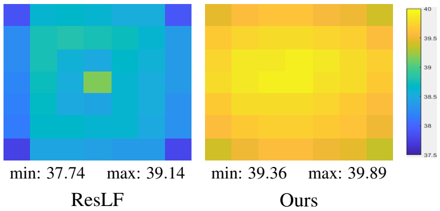


Figure 4. Comparison of the PSNR of the individual reconstructed view in *Bedroom*. The color of each grid represents the PSNR value.

Structure-aware loss function. The objective function is defined as the ℓ_1 error between the estimated LF image and the ground truth:

$$\ell_r = ||\widehat{L}^{rf} - L^{hr}||_1, \tag{7}$$

where $\widehat{L}^{rf}$ is the final reconstruction by the regularization module.

A high-quality LF reconstruction shall have strictly linear patterns on the EPIs. Therefore, to further enhance the parallax consistency, we add additional constraints on the output EPIs. Specifically, we incorporate the EPI gradient loss, which computes the ℓ_1 distance between the gradient of EPIs of our final output and the ground-truth LF, for the training of the regularization module. The gradients are computed along both spatial and angular dimensions on both horizontal and vertical EPIs:

$$\begin{aligned}\ell_e = &\|\nabla_x \widehat{E}_{y,v} - \nabla_x E_{y,v}\|_1 + \|\nabla_u \widehat{E}_{y,v} - \nabla_u E_{y,v}\|_1 \\ &+ \|\nabla_y \widehat{E}_{x,u} - \nabla_y E_{x,u}\|_1 + \|\nabla_v \widehat{E}_{x,u} - \nabla_v E_{x,u}\|_1,\end{aligned} \tag{8}$$

where $\widehat{E}_{y,v}$ and $\widehat{E}_{x,u}$ denote EPIs of the reconstructed LF images, and $E_{y,v}$ and $E_{x,u}$ denote EPIs of the ground-truth LF images.

4.3. Implementation and Training Details

Training strategy. To make the All-to-One SR module compatible for all different angular positions, we first trained it independently from the regularization network. During training, a training sample of an LF image was fed into the network, while a view at random angular posi-

Table 2. Quantitative comparisons (PSNR/SSIM) of different methods on 2× and 4× LF spatial SR. The best results are in bold, and the second best ones are underlined. PSNR/SSIM refers to the average value of all the scenes of a dataset.

		Bicubic	PCA-RR [8]	LFNet [28]	GB [23]	EDSR [19]	ResLF [37]	Ours
Stanford Lytro General [3]	2	35.93/0.940	36.44/0.946	37.06/0.952	36.84/0.956	39.34/0.967	40.44/0.973	**42.00/0.979**
Stanford Lytro Occlusions [3]	2	35.21/0.939	35.56/0.942	36.48/0.953	36.03/0.947	39.44/0.970	40.43/0.973	**41.92/0.979**
Kalantari *et al*. [14]	2	37.51/0.960	38.29/0.964	38.80/0.969	39.33/0.976	41.55/0.980	42.95/0.984	**44.02/0.987**
HCI new [12]	2	33.08/0.893	32.84/0.883	33.78/0.904	35.27/0.941	36.15/0.931	36.96/0.946	**38.52/0.959**
Inria Synthetic [24]	2	33.20/0.913	32.14/0.885	33.90/0.921	35.78/0.947	37.57/0.947	37.48/0.953	**39.53/0.963**
Stanford Lytro General [3]	4	30.84/0.830	31.24/0.841	31.30/0.844	30.38/0.841	33.15/0.882	33.68/0.894	**34.99/0.917**
Stanford Lytro Occlusions [3]	4	29.33/0.794	29.89/0.813	29.81/0.813	30.18/0.855	31.93/0.860	32.48/0.873	**33.86/0.895**
Kalantari *et al*. [14]	4	31.63/0.864	32.57/0.882	32.14/0.879	31.86/0.892	34.59/0.916	35.55/0.930	**36.90/0.946**
HCI new [12]	4	28.93/0.760	29.29/0.776	29.31/0.773	28.98/0.789	31.12/0.819	31.38/0.838	**32.27/0.859**
Inria Synthetic [24]	4	28.45/0.795	28.71/0.792	28.91/0.809	29.12/0.836	31.68/0.865	31.62/0.872	**32.72/0.890**

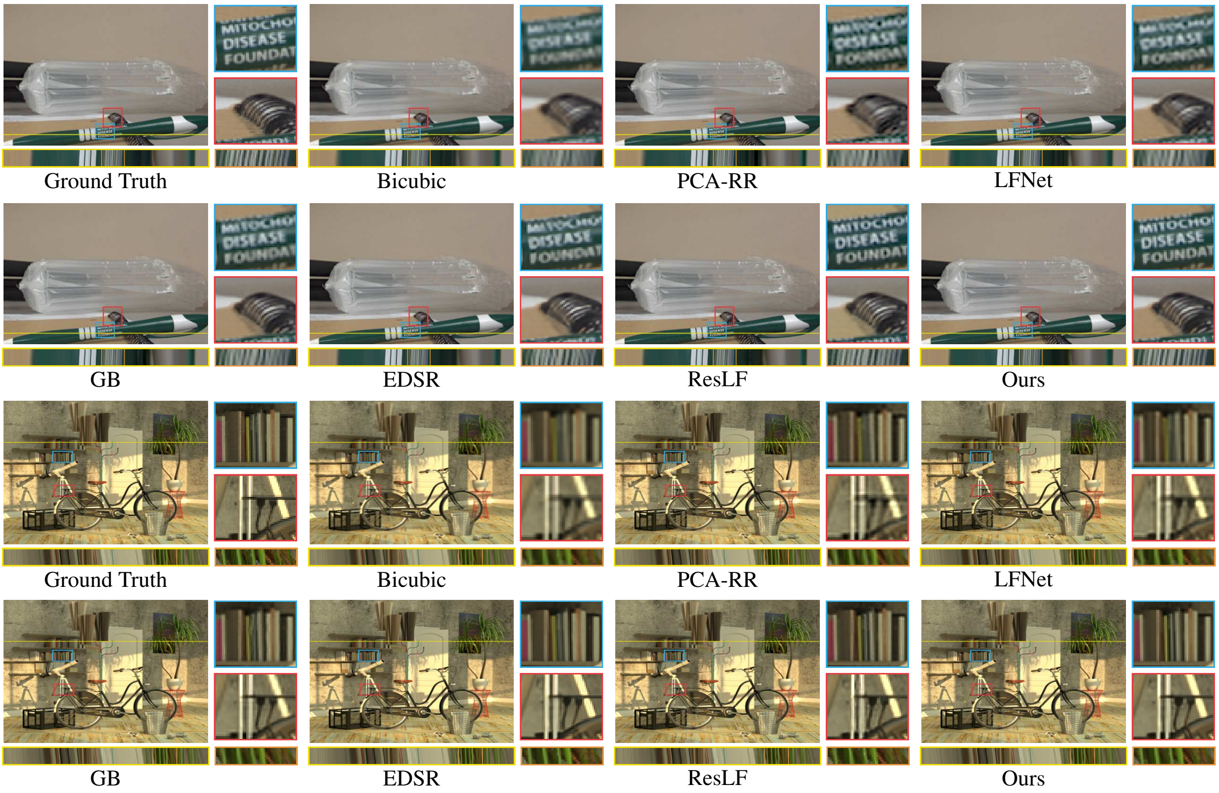


Figure 5. Visual comparisons of different methods on 2× reconstruction. The predicted central views, the zoom-in of the framed patches, the EPIs at the colored lines, and the zoom-in of the EPI framed patches in EPI are provided. Zoom in the figure for better viewing.

tion was selected as the reference view. After the All-to-One SR network training was complete, we fixed its parameters and used them to generate the intermediate inputs for the training of the subsequent structural consistency regularization module. The code is available at https://github.com/jingjin25/LFSSR-ATO.

Parameter setting. In our network, each convolutional layer has 64 filters with kernel size 3 × 3, and zero-padding was applied to keep the spatial resolution unchanged. In the per-view SR module, we set $n_1 = 5$, $n_2 = 2$, $n_3 = 2$ and $n_4 = 3$ for the number of residual blocks. For structural consistency regularization, we used $n_5 = 3$ alternate convolutional layers.

During training, we used LF images with angular resolution of 7 × 7, and randomly cropped LF patches with spatial size 64 × 64. The batch size was set to 1. Adam optimizer [17] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used. The learning rate was initially set to $1e^{-4}$ and decreased by

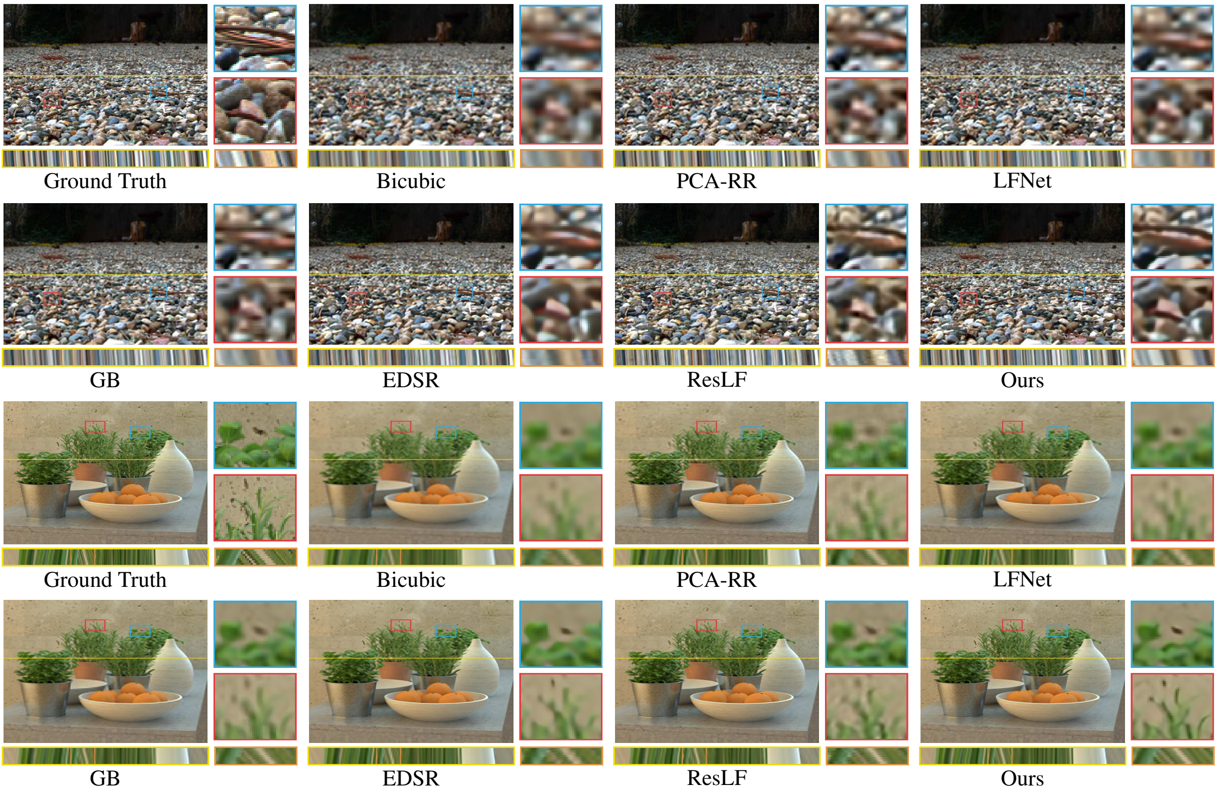


Figure 6. Visual comparisons of different methods on 4× reconstruction. The predicted central views, the zoom-in of the framed patches, the EPIs at the colored lines, and the zoom-in of the EPI framed patches in EPI are provided. Zoom in the figure for better viewing.

a factor of 0.5 every 250 epochs.

Datasets. Both synthetic and real-world LF datasets were used for training (180 LF images in total, which includes 160 images from Stanford Lytro LF Archive [3] and Kalantari *e.g*. [14], and 20 synthetic images from HCI [12]).

5. Experimental Results

4 LF datasets containing totally 138 real-world scenes and 43 synthetic scenes were used for evaluation. Details of the datasets and categories were listed in Table 4.2. Only Y channel was used for training and testing, while Cb and Cr channels were upsampled using bicubic interpolation when generating visual results.

5.1. Comparison with State-of-the-art Methods

We compared with 4 state-of-the-art LF SR methods, including 1 optimization-based method, i.e., GB [23], 3 learning-based methods, i.e., PCA-RR [8], LFNet [28], and ResLF [37], and 1 advanced SISR method EDSR [19]. Bicubic interpolation was evaluated as baselines.

Quantitative comparisons of reconstruction quality. PSNR and SSIM are used as the quantitative indicators for comparisons, and the average PSNR/SSIM over different testing datasets were listed in Table 2. It can be seen that our method outperforms the second best method, i.e. ResLF, by around 1 - 2 dB on both 2× and 4× SR.

We also compared the PSNR of individual views between ResLF [37] and ours, as shown in Figure 4. It can be observed that the gap between the central and corner views of our method is much smaller than that of ResLF. The significant degradation of the corner views in ResLF is caused by decreasing the number of views used for constructing directional stacks. Our method avoids this problem by utilizing the information of all views. Therefore, the performance degradation is greatly alleviated.

Qualitative comparisons. We also provided visual comparisons of different methods, as shown in Fig. 5 for 2× SR and Fig. 6 for 4× SR. It can be observed that most high-frequency details are lost in the reconstruction results of some methods, including PCA-RR, LFNet and GB. Although EDSR and ResLF could generate better results, some extent of blurring effects occurs in texture regions, such as the characters in the pen, the branches on the ground and the digits on the clock. In contrast, our method can produce SR results with sharper textures closer to the ground truth ones, which demonstrates higher reconstruction quality.

Comparisons of the LF parallax structure. As we dis-

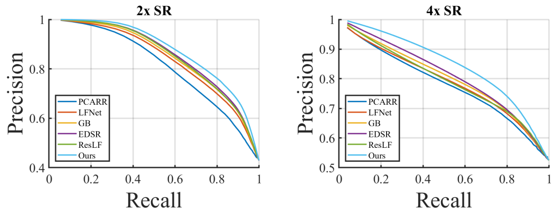


Figure 7. Quantitative comparisons of the LF parallax structure of the reconstruction results of different methods via LF edge parallax PR curves. The closer to the top-right corner the lines are, the better the performance is.

Table 3. Investigation of the effectiveness of the structural consistency regularization. The comparisons of the reconstruction quality before and after the regularization are listed. The top two rows show the average PSNR/SSIM over three datasets, and the rest rows show the comparisons on several LF images.

	w/o regularization	w/ regularization
Stanford Lytro Occlusions	41.62/0.978	41.92/0.979
HCI new	38.24/0.956	38.52/0.959
Occlusion_43_eslf	33.87/0.986	34.53/0.987
Occlusions_51_eslf	42.14/0.987	42.60/0.988
Antiques_dense	46.31/0.987	46.83/0.988
Blue_room_dense	40.19/0.978	40.66/0.979
Coffee_time_dense	35.09/0.977	35.65/0.979
Rooster_clock_dense	42.45/0.979	42.90/0.981
Cars	38.89/0.986	39.33/0.987
IMG_1554_eslf	37.02/0.989	37.46/0.991

Table 4. Comparisons of running time (in second) of different methods.

	Bicubic	PCA-RR [8]	GB [23]	EDSR [19]	ResLF [37]	Ours
2×	1.45	91.00	17210.00	0.025	8.98	23.35
4×	1.43	89.98	16526.00	0.024	8.79	7.43

cussed in Sec. 3, the straight lines in EPIs provide direct representation for the LF parallax structure. To compare the ability to preserve the LF parallax structure, the EPIs constructed from the reconstructions of different methods were depicted in Fig. 5 and Fig. 6. It can be seen that the EPIs from our methods show clearer and more consistent straight lines compared with those from other methods.

Moreover, to quantitatively compare the structural consistency, we computed the light filed edge parallax precision-recall (PR) curves [5], and Fig. 7 shows the results. The PR curves of the reconstructions by our method are closer to the top-right corner, which demonstrates the advantage of our method on structural consistency.

Efficiency comparisons. We compared the running time of different methods, and Table 4 lists the results of 4× reconstruction. Among them, learning-based methods were accelerated by a GeForce RTX 2080 Ti GPU. SISR methods, i.e., EDSR and bicubic, are faster than other compared methods, as all views can be processed in parallel. Although our method and ResLF are slightly slower than these SISR methods, much higher reconstruction quality is provided.

Table 5. Comparisons of the intuitive *All-to-All* fusion strategy and our *All-to-One*.

	All-to-All (image)	*All-to-All* (feature)	Ours *All-to-One*
General	41.25/0.977	41.25/0.977	**41.81/0.979**
Kalantari	43.18/0.985	43.13/0.985	**43.79/0.987**
HCI new	37.12/0.946	37.04/0.946	**38.24/0.956**

5.2. Ablation Study

***All-to-One* vs. *All-to-All*.** We compared the reconstruction quality of our proposed *All-to-One* fusion strategy with the intuitive *All-to-All* one, which simultaneously super-resolves all views by stacking the images or features of all views as inputs to a deep network. These networks were set to contain the same number of parameters for fair comparisons. Table 5 lists the results, where it can be seen that our *All-to-One* improves the PSNR by more than 0.6 dB on real-world data and 1.0 db on synthetic data, respectively, validating its effectiveness and advantage.

Effectiveness of the structural consistency regularization. We compared the reconstruction quality of the intermediate (before regularization) and final results (after regularization), as listed in Table 3. It can be observed that around 0.2-0.3 dB improvement is achieved on average over various datasets. For certain scenes, the contribution of the regularization is more obvious, such as 'Occlusion_43_eslf' and 'Antiques_dense', which obtain more than 0.5dB improvement by the regularization.

6. Conclusion and Future Work

We have presented a learning-based method for LF spatial SR. We focused on addressing two crucial problems, which we believe are paramount for high-quality LF spatial SR, i.e., how to fully take advantage of the complementary information among views, and how to preserve the LF parallax structure in the reconstruction. By modeling them with two sub-networks, i.e., All-to-One SR via combinatorial geometry embedding and structural consistency regularization, our method efficiently generates super-resolved LF images with higher PSNR/SSIM and better LF structure, compared with the state-of-the-art methods.

In our future work, other loss functions, such as the adversarial loss and the perceptual loss which have proven to promote realistic textures in SISR, and their extension to high-dimensional data can be exploited in LF processing.

References

[1] Lytro illum. https://www.lytro.com/. [Online]. 1

[2] Raytrix. https://www.raytrix.de/. [Online]. 1

[3] Raj Shah and Gordon Wetzstein Abhilash Sunder Raj, Michael Lowney. Stanford lytro light field archive. http://lightfields.stanford.edu/LF2016.html. [Online]. 5, 6, 7

[4] 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, 2012. 2

[5] Jie Chen, Junhui Hou, and Lap-Pui Chau. Light field denoising via anisotropic parallax analysis in a cnn framework. *IEEE Signal Processing Letters*, 25(9):1403–1407, 2018. 8

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 2

[7] Reuben Farrugia and Christine Guillemot. Light field super-resolution using a low-rank prior and deep convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[8] 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, 6, 7, 8

[9] Juliet Fiss, Brian Curless, and Richard Szeliski. Refocusing plenoptic images using depth-adaptive splatting. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–9, 2014. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 4

[11] Stefan Heber and Thomas Pock. Shape from light field meets robust pca. In *European Conference on Computer Vision (ECCV)*, pages 751–767, 2014. 3

[12] 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. 5, 6, 7

[13] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Transactions on Graphics*, 34(4):60, 2015. 1

[14] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*, 35(6):193:1–193:10, 2016. 5, 6, 7

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

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 3, 4, 5

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 624–632, 2017. 2

[19] 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. 6, 7, 8

[20] Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 22–28, 2012. 1, 2, 3

[21] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *IEEE International Conference on Computer Vision (ICCV)*, pages 261–270, 2017. 5

[22] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36, 2003. 2

[23] Mattia Rossi and Pascal Frossard. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Transactions on Image Processing*, 27(9):4207–4218, 2018. 1, 2, 3, 6, 7, 8

[24] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot. A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing*, pages 1–15, 2019. 5, 6

[25] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 5

[26] 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. 2

[27] Lipeng Si and Qing Wang. Dense depth-map estimation and geometry inference from light fields via global optimization. In *Asian Conference on Computer Vision (ACCV)*, pages 83–98, 2016. 1

[28] 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, 2, 3, 6, 7

[29] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *arXiv preprint arXiv:1902.06068*, 2019. 2

[30] Sven Wanner and Bastian Goldluecke. Spatial and angular variational super-resolution of 4d light fields. In *European Conference on Computer Vision (ECCV)*, pages 608–621, 2012. 2, 3

[31] 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, 2014. 1, 2, 3

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

[33] Wing Fung Henry Yeung, Junhui Hou, Jie Chen, Yuk Ying Chung, and Xiaoming Chen. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *European Conference on Computer Vision (ECCV)*, pages 137–152, 2018. 5

[34] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24(6):848–852, 2017. 1, 2, 3

[35] 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, 3

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

[37] 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. 1, 2, 3, 6, 7, 8

[38] Hao Zhu, Qing Wang, and Jingyi Yu. Light field imaging: models, calibrations, reconstructions, and applications. *Frontiers of Information Technology & Electronic Engineering*, 18(9):1236–1249, 2017. 1

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

[40] Hao Zhu, Qi Zhang, and Qing Wang. 4d light field superpixel and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6384–6392, 2017. 1