Residual Networks for Light Field Image Super-Resolution

Shuo Zhang[1,2] , Youfang Lin[1,2], Hao Sheng[3,4]

[1]Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University

[2]Key Laboratory of Intelligent Passenger Service of Civil Aviation, CAAC

[3]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University

[4]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University

{zhangshuo, yflin}@bjtu.edu.cn, shenghao@buaa.edu.cn

Abstract

Light field cameras are considered to have many potential applications since angular and spatial information is captured simultaneously. However, the limited spatial resolution has brought lots of difficulties in developing related applications and becomes the main bottleneck of light field cameras. In this paper, a learning-based method using residual convolutional networks is proposed to reconstruct light fields with higher spatial resolution. The view images in one light field are first grouped into different image stacks with consistent sub-pixel offsets and fed into different network branches to implicitly learn inherent corresponding relations. The residual information in different spatial directions is then calculated from each branch and further integrated to supplement high-frequency details for the view image. Finally, a flexible solution is proposed to super-resolve entire light field images with various angular resolutions. Experimental results on synthetic and real-world datasets demonstrate that the proposed method outperforms other state-of-the-art methods by a large margin in both visual and numerical evaluations. Furthermore, the proposed method shows good performances in preserving the inherent epipolar property in light field images.

1. Introduction

With recent advances in camera devices, light field (LF) imaging technology is commonly used in the market for 3D reconstruction and virtual reality applications [12, 28, 14]. For large-scale applications, the camera array is often used to capture high-resolution LF images with a large baseline. By inserting the micro-lens array between the main lens and the imaging plane [1], the handheld plenoptic camera [12, 14] is developed and is able to capture LF images with a small baseline by one shot, which has more broad applications such as image refocusing [13]. However, the development of plenoptic cameras is severely limited due to their lower spatial resolutions compared with traditional cameras, which has brought a lot of difficulties in many practical vision applications.

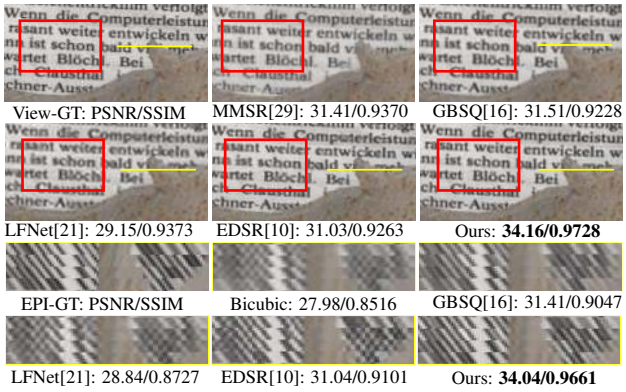


Figure 1. The spatially super-resolution results of image *Horses* [25] with $\times 2$ magnification factor. Our results of the central view images and epipolar plane images (EPIs) outperform the other state-of-the-art methods with significant higher PSNR and SSIM. The background letters along the occlusion boundary are clearly recovered with sharp edges both in view and epipolar plane images using the proposed method, while the others exhibit strong artifacts or ambiguous textures.

Since light fields record scenes with multiple view images, disparity information in these view images provides multiple sampling with sub-pixel offsets to enhance the spatial resolution. Traditional methods register the sub-pixel information by explicitly warping other view images based on prior disparity information [16, 23, 29]. However, existing disparity estimation methods for LF images suffer from occlusions, noises and textureless regions [7], which lead to significant artifacts in reconstructed LF images. Recently, deep-learning-based methods have been proposed for light field super-resolution (LFSR), in which disparity information is implicitly learned during training processes [21, 27]. However, these methods are quite limited in exploring ac-

curate sub-pixel information and preserving the inherent epipolar property in LF images.

Taking advantage of the residual structure in super-resolution networks [8, 10, 30], we design a novel residual network (resLF) to enhance the spatial resolution of LF images. In the proposed method, the view images in one LF are first separated into four groups according to their angular directions and fed into different network branches to learn high-frequency details in the specific spatial directions. Different from other LFSR methods, inherent corresponding relations in view images, which reflect disparity information, are implicitly explored and sub-pixel mappings from various directions are learned in the proposed method. Residual information from different spatial directions is then combined to generate complete residual details for final super-resolved central view images. The LF is divided into different parts and the entire view images are finally super-resolved based on a flexible solution.

The experiments are conducted on different LF images and various challenging scenes, which include noises, occlusions and non-Lambertian surfaces. The resLF networks can be used for both synthetical and real-world LF images with different angular resolutions. Experimental results show that the proposed framework significantly outperforms the other state-of-the-art methods in terms of numerical and visual evaluations, where the PSNR results are improved by 1.5 dB on average in $\times 2$ and $\times 4$ super-resolution tasks. Moreover, the comparison of epipolar plane images (EPIs) shows that the proposed method is able to preserve the corresponding relations in super-resolved view images.

2. Related Work

As LFs capture scenes with multiple view images from different angles, texture information lost in spatial domain actually remains in angular domain according to scene structures. Most LFSR methods can be divided into two categories, disparity-based and learning-based methods, based on learning structure information directly or implicitly.

Given estimated structure information as priors, many researches focus on how to warp multiple view images accurately and find sub-pixel information to improve the spatial resolution. Based on EPIs, Wanner *et al.* [23, 24] extracted depth information using structure tensor and interpolated lines in EPIs accordingly to super-resolve view images. By modeling LF patches based on a Gaussian Mixture Model, Mitra *et al.* [11] proposed to reconstruct patches with higher resolutions based on estimated disparities. The other methods [2, 3] focused on recovering view images by explicitly warping pixels from other view images. Recently, Zhang *et al.* [29] proposed to estimate matching relationships between micro-lens images and view images, and used micro-lens images with richer textures to recover

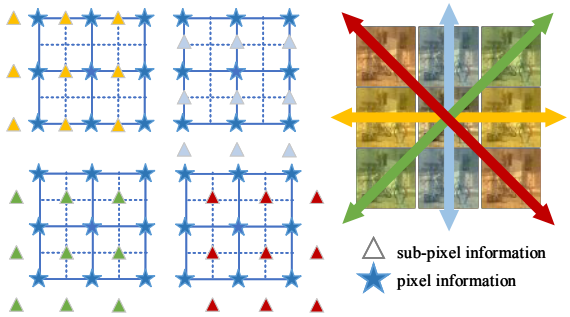


Figure 2. Sub-pixel information from surrounding view images in LF. Different images contain sub-pixel shifts in different directions according to the disparity information, *e.g.*, the horizontal sub-pixel position (labeled as yellow) can be found in the horizontal adjacent images. In our network, we propose to learn the mapping from sub-pixel shifts in surrounding image to a high-resolution central view image.

related view images. Based on a graph-based regularizer, Rossi *et al.* [16] designed a global optimization problem to augment the resolution of all LF view images together. The disparity of each view is roughly estimated to calculate the warping matrix and the geometric structure between each view is considered to optimize the super-resolution results. However, although lots of disparity estimation methods have been proposed [7], reconstructed view images are still easily affected by estimation errors, which cause significant artifacts along occlusion boundaries.

Recently, deep Convolutional Neural Networks (CNN) have been developed for single image super-resolution (SISR) [4] and achieved remarkable performances by introducing residual learning, recursive layers or deeper CNN models [8, 20, 30]. For LF images, several frameworks are designed to implicitly learn geometric structures and augment the spatial resolution. Cho *et al.* [3] proposed to train a dictionary for high and low quality LF image pairs and improved the image quality based on sparse coding. Yoon *et al.* [26, 27] developed spatial and angular networks to upsample angular and spatial resolution simultaneously. The different view images are augmented separately and then combined into different types of image pairs to create novel views. Wang *et al.* [21] built a bidirectional recurrent CNN to super-resolve horizontal and vertical image stacks separately and then combined them using stacked generalization technique to obtain complete view images. In their method, the spatial relations are iteratively investigated between two adjacent view images. However, as most learning-based frameworks chose to super-resolve each view image pair or stack separately, corresponding relations in different views are not fully considered so that accurate high-frequency details cannot be well recovered.

Since view images in LF capture scenes from various directions, high-frequency details in spatial domain are actu-

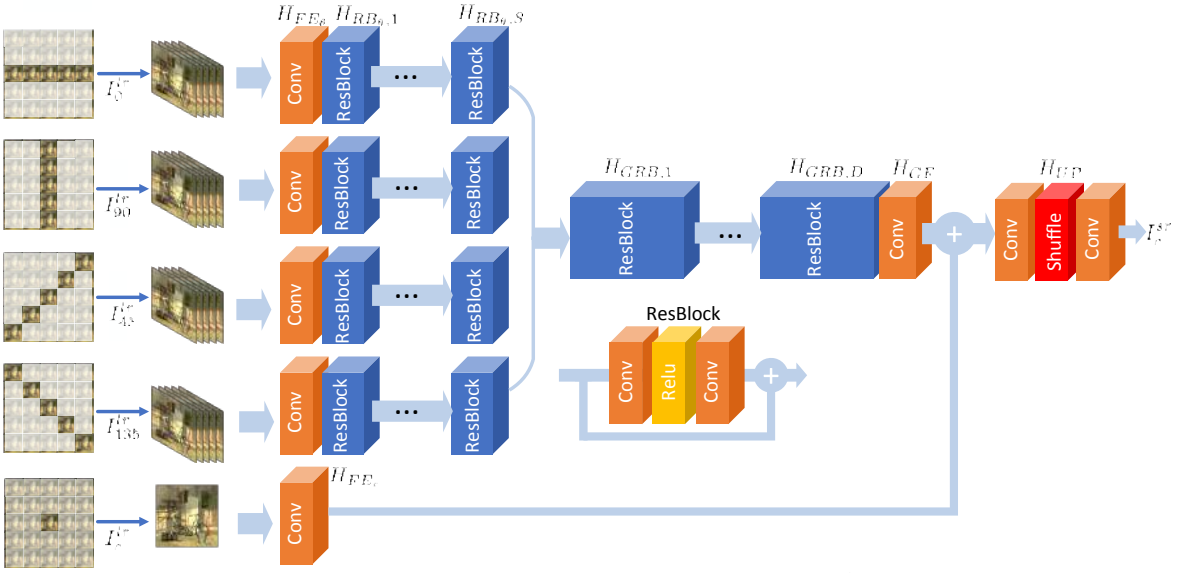


Figure 3. The overall structure of the proposed resLF network. The different image stacks I_θ^{lr} are fed into different network branches including feature extraction layer (H_{FE_θ}) and S residual blocks ($H_{RB_{\theta,s}}$). The output of each branch is then concatenated for D global residual blocks ($H_{GRB_\theta,d}$). The central image I_c^{lr} after feature extraction (H_{FE_c}) is then added to the global residual output. Finally, the up-sampling network (H_{UP}) is introduced to obtain final super-resolution results.

ally kept in view images from different angular directions, as shown in Fig 2. This special architecture provides the possibilities of finding sub-pixel information in spatial domain from angular domain. Different from the above methods, we propose a specifically designed super-resolution framework to find accurate sub-pixel information from different angular directions and preserve the epipolar property at the same time.

3. Methodology

The objective of the proposed resLF network is to reconstruct a super-resolution (SR) LF image $L^{sr}(x, y, u, v)$ from a low-resolution (LR) image $L^{lr}(x, y, u, v)$, where (x, y) is in the spatial domain and (u, v) is in the angular domain [9]. Assuming that the resolution of L^{lr} is described with (X, Y, U, V), L^{sr} with higher spatial resolution can be denoted with (rX, rY, U, V), where r represents the up-sampling factor in spatial resolution and $U = V$ in most LF images. We convert images to YCbCr color space and only deal with Y channel images.

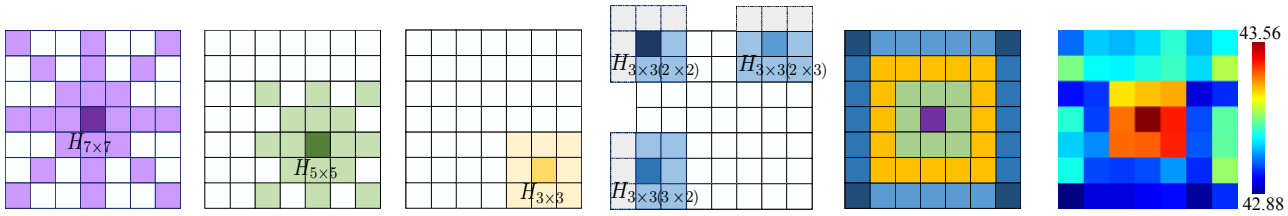
3.1. Framework Overview

The view images in one LF, *i.e.* sub-aperture images, capture scenes in different directions, which can be extracted by fixing (u, v) and changing (x, y) coordinates. Different from traditional multi-view images, viewpoints in LFs have various angles. As shown in Fig.2, the surrounding view image in one angular direction contains sub-pixel offsets in the specific spatial direction. The shifted pixels in different directions can be combined according to disparity information to yield the high-resolution view image. Therefore, we propose to explore the detail information from surrounding view images which have horizontal, vertical or diagonal sub-pixel shifts.

If angular direction $\tan\theta = v/u$, we extract image stacks $I_{\theta=0}$, $I_{\theta=90}$, $I_{\theta=45}$, $I_{\theta=135}$ around one view image, whose viewpoints change along horizontal, vertical, left and right diagonal directions. Inspired by the great performance of the recently residual learning in SISR [30], we specifically design a residual network structure for LF images. As shown in Fig. 3, the resLF network contains four branches and a global part with several residual blocks. Compared with other view images, the central view image in one LF has more available sub-pixel information from the related image stacks. Therefore, we first design the network to improve the spatial resolution of the central view image (see Sec. 3.2), where the number of available view images in each image stack is the same. The network is then trained using LF images with different angular resolutions and different image stack inputs. Finally, the entire LF images are recovered according to a flexible solution (see Sec. 3.3).

3.2. Network Design

Suppose that the training data $\{L^{lr}, I_c^{hr}\}$ is given. I_c^{hr} is the ground truth, which represents the central view image with high resolution. The four image stacks from different directions $\{I_0^{lr}, I_{45}^{lr}, I_{90}^{lr}, I_{135}^{lr}\}$ around the central view image I_c^{lr} can be calculated. The objective of our network is to learn a model $H_{U\times V}$ that can predict a high-resolution cen-



(a) $H_{7\times7}$ (b) $H_{5\times5}$ (c) $H_{3\times3}$ (d) $H_{3\times3(2\times3,3\times2,2\times2)}$ (e) Final LF images (f) PSNR values

Figure 4. The flexible solution for super-resolving complete LF images. As shown in (a–d), the LF is divided into different parts and each central view image in the LF part is super-resolved using corresponding networks. (e) shows all the views in one LF where each color represents the corresponding network. In (f), the color of each grid represents the PSNR of the reconstructed view image in *Mona*.

tral view image $I_c^{sr} = H_{U\times V}(I_0^{lr}, I_{45}^{lr}, I_{90}^{lr}, I_{135}^{lr}, I_c^{lr})$ from given input with $U \times V$ angular resolution.

Considering the narrow baseline in adjacent view images, the image stack from one direction is directly concatenated as the input of each branch. Similar with [18], the network is constructed with four branches and each image stack is encoded individually to learn the residual part in the specific direction. The disparity information is calculated implicitly using convolutional networks to find out accurate sub-pixel shifts between each view image. In each branch, the first convolutional layer $H_{FE_\theta}(\cdot)$ extracts features from each image stack:

$$F_{FE_\theta} = H_{FE_\theta}(I_\theta^{lr}), \tag{1}$$

where I_θ^{lr} denotes the image stack in each direction θ and F_{FE_θ} is given as the input of the following residual blocks. We also define a similar convolution operation $H_{FE_c}(\cdot)$ for the central view image I_c^{lr} to extract corresponding features for residual learning:

$$F_{FE_c} = H_{FE_c}(I_c^{lr}). \tag{2}$$

Suppose that we have S residual blocks (RB) in each branch, the output $F_{RB_\theta,s}$ of the s-th RB can be calculated as:

$$F_{RB_\theta,s} = H_{RB_\theta,s}(F_{RB_\theta,s-1}), \tag{3}$$

where $H_{RB_\theta,s}(\cdot)$ denotes the s-th RB. The structure of RB is defined similar with [10], which contains a convolution layer, rectified linear units (ReLU) and a convolution layer in order. In each block, the input is directly added to the output as a residual part.

After extracting features from RBs with different directions, we integrate all the features in a global way and further feed them into more residual learning blocks:

$$\begin{aligned} F_{GRB,d} &= H_{GRB,d}(F_{GRB,d-1}) \\ &= H_{GRB,d}(\cdots H_{GRB,1}(F_{GRB,0})\cdots), \end{aligned}$$

where $F_{GRB,0} = [F_{RB_0,S}, F_{RB_{90},S}, F_{RB_{45},S}, F_{RB_{135},S}]$ refers to the concatenation of features from S residual blocks in each branch. Since the image stack is grouped according to different angular directions, the output of each group is corresponding to sub-pixels in the specific spatial direction. We define D global residual block $H_{GRB,d}(\cdot)$ to further exploit the sub-pixel residual information from different directions. $H_{GRB,d}(\cdot)$ has the similar structure with residual block $H_{RB,s}(\cdot)$ in each branch but has 4 times larger number of filters to extract more features.

The global features are then fed into a convolution layer $H_{GF}(\cdot)$ and combined with the features F_{FE_c} from the central view image. After extracting local and global features for the central image in the low-resolution space, we introduce an up-sampling net $H_{UP}(\cdot)$ to obtain the final image in the high-resolution space. Inspired by the work in [17], one convolution layer and one shuffle layer followed by one convolutional layer is used to construct the up-sampling net. The final super-resolved view image can be denoted as:

$$I_c^{sr} = H_{UP}(H_{GF}(F_{GRB,D}) + F_{FE_c}), \tag{4}$$

where $F_{GRB,D}$ refers to the output of D global residual blocks. The super-resolved view image combines the residual information learning from surrounding view images with different directions.

3.3. Light Field Super-Resolution

As different view images of one LF capture scenes in their specific directions and have their specific features, it is difficult to super-resolve all view images in one network simultaneously. State-of-the-art learning-based LFSR algorithms built complex networks to reconstruct complete LF images by either applying the super-resolution process for each image individually [27] or calculating view images based on already super-resolved view images [21]. In this way, the view images in one LF are super-resolved with unbalanced information and related results show big differences. At the same time, corresponding relations in view images are also hard to preserve. Moreover, for LF images with different angular resolutions, these networks should be trained from the beginning, whose performances decrease sharply for LFs with small angular resolution. Different from these complicated networks, we choose to deal with each view image individually by combining surrounding

Table 1. Quantitative comparisons using different network structures for $\times 2$ super-resolution results.

Network	Incomplete Images Stacks			Different Angular Resolutions				Different Image Stacks		
	$H_{3\times3(3\times2)}$	$H_{3\times3(2\times3)}$	$H_{3\times3(2\times2)}$	$H_{3\times3}$	$H_{5\times5}$	$H_{7\times7}$	$H_{9\times9}$	$H_{1\times9,h}$	$H_{9\times1,v}$	$H_{9\times9,c}$
Avg. PSNR	37.19	37.09	37.02	37.24	37.60	37.65	**37.77**	36.27	36.32	37.35
Avg. SSIM	0.9756	0.9748	0.9742	0.9756	0.9777	0.9782	**0.9788**	0.9687	0.9701	0.9768

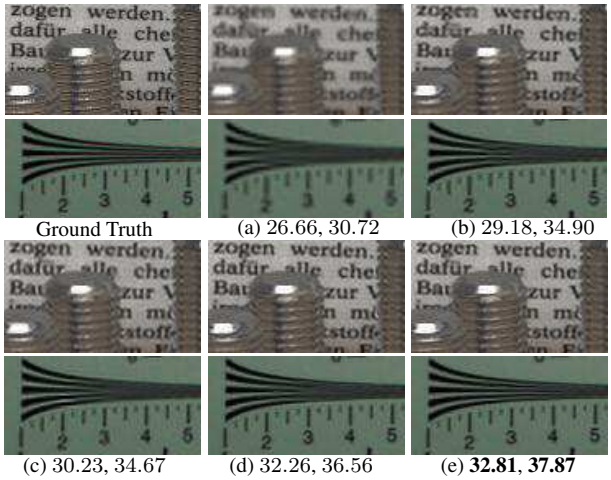
view images to keep geometric structures, which also provides a more flexible solution to obtain entire LF images with various angular resolutions.

In order to super-resolve complete LF images, we train different resLF networks with various angular resolutions. The LF image is divided into different parts, where the other view images are treated as the central view image in the corresponding LF part. As in Fig. 4, we show the super-resolution process for the LF image with 7×7 angular resolution. For the central view image, each stack has 7 images and network $H_{7\times7}$, is used to generate the result. The adjacent view images are recovered using $H_{5\times5}$ and $H_{3\times3}$ network, respectively. As for border view images, we lack the image information from one or more directions in the image stacks. One solution is to change the network structure with different branches for border views using available surrounding views. In the proposed framework, we choose another solution to pad lacking views with 0 in each image stack and use the same network structure. Specifically, we train network $H_{3\times3(3\times2)}$ for left and right border images, $H_{3\times3(2\times3)}$ for up and down border images and $H_{3\times3(2\times2)}$ for four corner images, separately. The resLF networks with various angular resolutions are then used for super-resolving all view images in LFs.

We evaluate the different resLF networks on a part of our test dataset and the results are shown in Table 1. The network with more view images in each image stack achieves better performances in general. For LF images with 3×3 angular resolution, even for border images with incomplete image stacks, the proposed network also produces comparable super-resolution results after training accordingly. The numerical results for the example image *Mona* are shown in Fig. 4. For this image, the center view image achieves the highest accuracy and the border view images also obtain similar performances.

3.4. Implementation Detail

Due to the small baseline of LF images, we set 3×3 as the kernel size and pad zero in all convolutional layers. The convolutional layers in the individual branch have 32 filters and the layers in global residual blocks after concatenation have 32×4 filters. The global feature extraction layer has 32 filters to combine the residual details with the original image. The final convolutional layer has 1 channel to output the desired SR image. We set $S = 4$ in each branch and



Ground Truth (a) 26.66, 30.72 (b) 29.18, 34.90
(c) 30.23, 34.67 (d) 32.26, 36.56 (e) **32.81**, **37.87**

Figure 5. Super-resolution results using different network structures, where the PSNR value of each image is shown accordingly. (a) Bicubic Interpolation (b)$H_{9\times1,v}$ (c)$H_{1\times9,h}$(d) $H_{9\times9,c}$ (e) $H_{9\times9}$. The network combined with image stacks from more directions obtains sharper edges and textures than the others.

$D = 4$ in the global part after several experimental tests. The $L1$ loss function is used as it provides better performances than $L2$ loss function in the proposed network.

In each training batch, we randomly extract 64 LF patches with the spatial size of 35×35 as inputs. We randomly augment the dataset by flipping the images horizontally or vertically, or rotating 180 degree. As pixel shifts are related to disparity information, the image order in each stack should be changed accordingly to preserve the epipolar property in LF images. We train our model with Adam optimizer and the weights in each layer are initialized using Xaviers algorithm [5]. The learning rate is initialized to 10^{-3} for all weights and decreases by a factor of 0.1 for every 100 epoch. We implement the resLF network with Torch7 framework and the training process roughly takes 1 day with a Titan GPU. Using the proposed model, each view image can be spatially $\times 2$ super-resolved within 7 ms.

4. Experiment

The synthetic images from HCI1 [25] and HCI2 [6], and real-world images [19, 15, 21] from Lytro Illum cameras [12] are used in the experiment. The training, validation and test datasets are chosen from the above im-

Table 2. Quantitative evaluations (PSNR / SSIM) of $\times 2$ super-resolution results on synthetic light field image *Buddha* and *Mona* .

Methods	*Buddha*			*Mona*		
	Min	Avg	Max	Min	Avg	Max
Mitra [11]	28.83 / 0.8665	29.91 / 0.8994	31.17 / 0.9343	28.54 / 0.8541	29.28 / 0.8911	30.07 / 0.9319
Wanner [23]	24.43 / 0.7662	29.69 / 0.8691	36.97 / 0.9470	25.40 / 0.8542	30.76 / 0.9324	37.60 / 0.9862
Yoon [27]	36.25 / 0.9579	36.95 / 0.9623	37.35 / 0.9657	37.03 / 0.9833	37.99 / 0.9863	38.53 / 0.9878
MMSR [29]	- / -	39.83 / 0.9745	- / -	- / -	34.44 / 0.9702	- / -
GBSQ [16]	39.29 / 0.9678	40.00 / 0.9730	40.37 / 0.9754	39.88 / 0.9795	40.41 / 0.9812	40.73 / 0.9821
LFNet [21]	38.09 / 0.9709	38.42 / 0.9731	38.77 / 0.9760	38.38 / 0.9884	38.73 / 0.9891	38.80 / 0.9895
EDSR [10]	39.72 / 0.9680	39.93 / 0.9703	40.43 / 0.9726	42.21 / 0.9800	42.35 / 0.9803	42.46 / 0.9806
Proposed	**41.09 / 0.9881**	**41.62 / 0.9897**	**42.24 / 0.9910**	**42.88 / 0.9929**	**43.13 / 0.9934**	**43.56 / 0.9941**

ages, which include 250, 50 and 50 LF images, respectively. All the dataset are preprocessed with 9×9 angular resolution and are cropped with smaller angular resolutions to train different networks. The images are spatially $\times 2$ and $\times 4$ downsampled using bicubic interpolation and super-resolved using the proposed method. We train the synthetic and real-world images all together to obtain robust super-resolution results. The super-resolution results are evaluated with PSNR and SSIM [22]. The results are compared with state-of-the-art LFSR algorithms, including disparity-based method [23, 11, 29, 16] and learning-based methods [27, 21]. The state-of-the-art SISR method [10] is also used for further comparisons.

4.1. Ablation Investigation

In this subsection, we investigate the performances of networks with different branches. We design network $H_{1\times 9,h}$ and $H_{9\times 1,v}$ using only horizontal or vertical image stacks as input, and $H_{9\times 9,c}$ using both horizontal and vertical image stacks. The number of parameters and the architectures of the networks are kept the same with the original network for a fair comparison.

The quantitive comparisons on the part of our test dataset are illustrated in Table 1. The network with image stacks from more angular directions achieves better performances than single angular direction. Moreover, compared with $H_{9\times 9,c}$, $H_{5\times 5}$ network with input images from more directions obtains better results, where the number of input view images in each network is almost the same. We also show some examples in Fig. 5, which contain complex textures along with a lot of occlusions. It is obvious that the results from $H_{9\times 9}$ network achieve better visual effects than the others.

4.2. EPIs Comparison

As we deal with each view using different networks, it is important to verify whether the method is able to keep the inherent geometric structure. We integrate the super-resolved views into EPIs and compare the epipolar property in Fig. 1, Fig. 6 and Table 3, where the SSIM values of EPIs are provided. As shown, EPIs from GBSQ [16] have blurry results. Although the image stacks are simultaneously super-resolved in LFNet [21], the oblique lines are still distorted since the epipolar constraint in each image stack is not fully considered in the super-resolution process. As EDSR [10] only focuses on one single view image, high-frequency textures are super-resolved individually in each view image so that the corresponding cues in EPIs are disordered. By contrast, our method uses surrounding view images to take the epipolar constraint into consideration for each view image so that the geometric structure is well preserved in final reconstructed LF images.

4.3. Sythetic Images

We compare the proposed method with state-of-the-art methods on synthetic LF images (HCI1 [25] and HCI2 [6]) in this subsection. The LF images are cropped with 5×5 angular resolution and super-resolved with $\times 2$ spatial resolution. As explained in Sec. 3.3, the networks $H_{5\times 5}, H_{3\times 3}, H_{3\times 3(3\times 2)}, H_{3\times 3(2\times 3)}$ and $H_{3\times 3(2\times 2)}$ are used to super-resolve different view images in one LF. The detailed comparisons for image *Buddha* and *Mona* are listed in Table 2, where the minimum, average and maximum PSNR and SSIM is shown. Results from disparity-based methods [23, 11, 29] vary greatly in different view images due to the uncertain disparity information. As most of the textures in *Mona* are regular, the EDSR [10] achieves the second best scores compared with the other LF specific methods. Our method outperforms the other methods with more than 1.62 dB (PSNR) in *Buddha* and 0.78 dB (PSNR) in *Mona*. The differences in our super-resolved view images are small for one LF and the minimum PSNR and SSIM values of our results are still higher than the others.

We provide the average results on the synthetic datasets in Table 3. The proposed resLF exceeds second best results by 1.62 dB in PSNR and 0.02 in SSIM on average. The qualitative comparisons are shown in Fig. 1 and Fig. 6. MMSR [29] obtains clear textures in flat regions but fails in occlusion edges due to the wrong estimated disparity information. EDSR [10] cannot predict complex textures

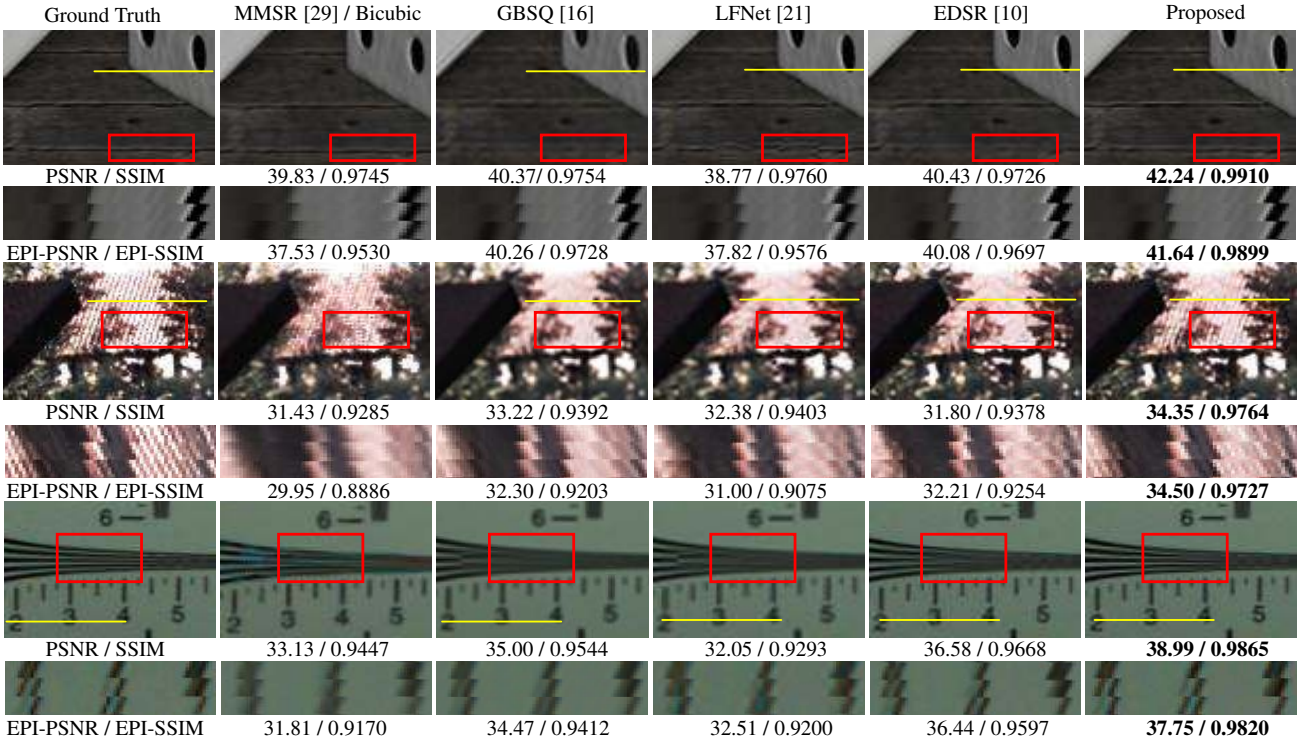


Figure 6. The detailed $\times 2$ super-resolution results for synthetic image *Buddha* [25] and real-world image *occlusion_4* [19], *ISO_Chart_12* [15]. The super-resolved central view images and EPIs are shown, where the corresponding PSNR and SSIM values are illustrated below. As MMSR [29] is only effective for central view in LF, we show the EPIs using Bicubic methods. Our method is able to recover more accurate details in view images and preserve the epipolar features in EPIs than the others.

Table 3. Avg. PSNR/SSIM of the $\times 2$ super-resolved view images in each LF dataset and Avg. SSIM of the related EPIs in all datasets.

Methods	HCI1 [25]	HCI2 [6]	Lytro [15]	Bikes [19]	Occlusions [19]	Reflective [19]	Overall EPIs
Bicubic	35.23 / 0.9303	31.67 / 0.8816	31.23 / 0.8856	29.76 / 0.9014	33.60 / 0.9273	36.94 / 0.9495	0.9210
MMSR [29]	35.44 / 0.9621	31.46 / 0.9189	29.83 / 0.9284	29.83 / 0.9284	33.38 / 0.9440	36.13 / 0.9571	-
GBSQ [16]	38.04 / 0.9635	34.61 / 0.9423	32.46 / 0.9295	31.69 / 0.9445	36.23 / 0.9596	38.29 / 0.9649	0.9411
LFNet [21]	36.46 / 0.9645	33.63 / 0.9317	32.70 / 0.9348	31.92 / 0.9499	35.92 / 0.9630	38.80 / 0.9706	0.9367
EDSR [10]	39.24 / 0.9657	35.07 / 0.9489	33.94 / 0.9473	33.86 / 0.9638	37.61 / 0.9692	40.64 / 0.9758	0.9560
Proposed	**41.09 / 0.9882**	**36.45 / 0.9786**	**35.48 / 0.9727**	**35.21 / 0.9806**	**39.71 / 0.9876**	**42.32 / 0.9904**	**0.9778**

based on one single view image and produces blurry details. LFNet [21] is trained with horizontal and vertical image stacks and results are combined in the stacked generalization. As they analyzed in [21], their final results after the combination only achieve tiny improvements in PSNR, which means sub-pixel information from different directions is not well integrated. Therefore, the related results are recovered with artifacts. In our method, the proposed network combines surrounding view images from different directions in a global way so that it is able to deal with complex textures. The results show significantly better image qualities where the textures are recovered accurately in both flat and occlusion regions.

We also train a set of $4\times$ networks for a harder super-resolution task. The quantitative results are illustrated in Table 4 and the qualitative results are shown in Fig. 7. Our method also outperforms the others with more than 1 dB in most of the evaluated images. The results from EDSR [10] are over-smoothed with ambiguous details and LFNet [21] produces blurry and noisy results. By contrast, our results preserve the boundaries and textures well and show superior performances in visual effects.

4.4. Real-World Images

As the proposed networks are designed and trained for all kinds of LF images, the models can be directly used for Lytro images. The $\times 2$ results are shown in Table 3 and Fig. 6, where different categories in [19] are evaluated separately. The $\times 4$ results are compared in Fig. 7 and Table 4. As these LF images are captured using plenoptic cameras, original view images contain noticeable artifacts and noises, which make it more difficult to estimate disparity in-

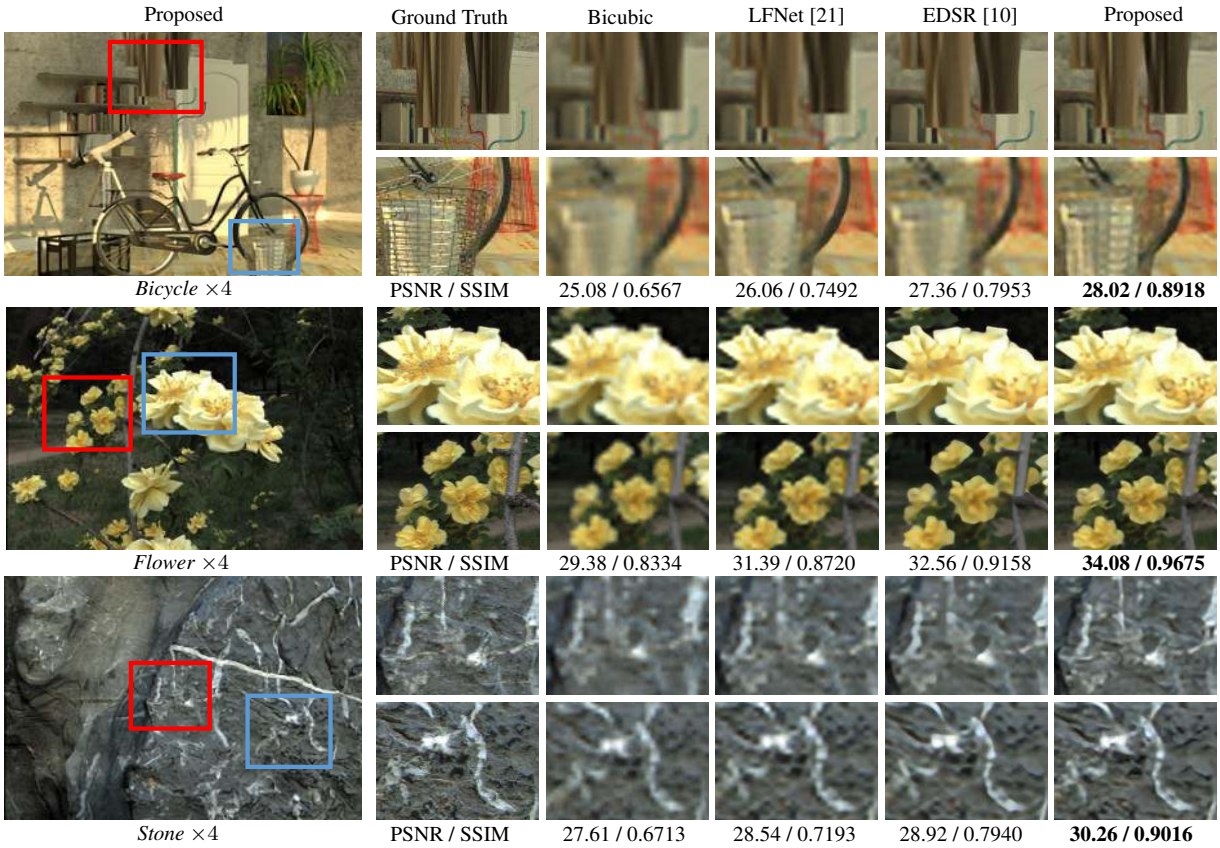


Figure 7. Image comparisons of $\times 4$ super-resolution results, where the super-resolved central view image is shown. The average PSNR and SSIM values are illustrated below. Our results show fewer artifacts and superior image quality compared with other methods.

Table 4. Quantitative evaluations (Avg. PSNR / Avg. SSIM) of $\times 4$ super-resolution results on different light field datasets.

Methods	*Budda* [25]	*Sideboard* [6]	*ISO_Chart_12* [15]	*Reeds* [15]	*Stone* [21]	*Flower* [21]
Bicubic	32.30 / 0.8470	23.41 / 0.5886	26.60 / 0.7755	37.09 / 0.8853	27.61 / 0.6713	29.38 / 0.8334
LFNet [21]	33.51 / 0.8827	24.67 / 0.7273	27.59 / 0.8776	37.71 / 0.9624	28.54 / 0.7193	31.39 / 0.8720
EDSR [10]	34.98 / 0.9059	26.10 / 0.7968	30.96 / 0.9148	38.02 / 0.9071	28.92 / 0.7940	32.56 / 0.9158
Proposed	**36.06 / 0.9623**	**27.35 / 0.8840**	**32.57 / 0.9551**	**39.25 / 0.9731**	**30.26 / 0.9016**	**34.07 / 0.9675**

formation and reconstruct LF images, especially for the $4\times$ task. The LFNet in [21], which trains real-world images especially, produces obvious artifacts in the super-resolved images. The results from EDSR [10] are ambiguous and over-smoothed since the information from other view images is not considered. By contrast, our results achieve significantly higher PSNR and SSIM in different kinds of real-world images for both tasks, which shows that the proposed network can not only handle the noisy input but also recover more high-frequency details.

5. Conclusions

In this paper, a residual convolutional network has been proposed to augment the spatial resolution of light field images. The inherent structure information in light field images is explored in different network branches from different angular directions and used to infer sub-pixel information in high-resolution view images through the network. The entire light field images with different angular resolutions can be super-resolved based on different trained models. Experimental results show that our method outperforms the state-of-the-art methods by a large margin in PSNR and SSIM and exhibits significantly better visual effects. The proposed network can preserve the epipolar property of the images well and can be used for different kinds of light field images with different angular resolutions.

Acknowledgment: This work is supported by the Fundamental Research Funds for the Central Universities 2019RC013.

References

[1] Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):99–106, 1992.

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

[3] Donghyeon Cho, Minhaeng Lee, Sunyeong Kim, and Yu-Wing Tai. Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3280–3287, 2013.

[4] 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.

[5] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

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

[7] Ole Johannsen, Katrin Honauer, Bastian Goldluecke, Anna Alperovich, Federica Battisti, Yunsu Bok, Michele Brizzi, Marco Carli, Gyeongmin Choe, Maximilian Diebold, et al. A taxonomy and evaluation of dense light field depth estimation algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1795–1812, 2017.

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

[9] Marc Levoy and Pat Hanrahan. Light field rendering. In *ACM Proceedings of Annual Conference on Computer Graphics and Interactive Techniques*, pages 31–42, 1996.

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 4, 2017.

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

[12] Ren Ng. Lytro redefines photography with light field cameras. `http://www.lytro.com`. Accessed: Oct. 22, 2018.

[13] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11), 2005.

[14] Christian Perwa and Lennart Wietzke. Raytrix: Light filed technology. `http://www.raytrix.de`. Accessed: Oct. 22, 2018.

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

[16] 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.

[17] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

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

[19] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. The stanford lytro light field archive. `http://lightfields.stanford.edu/LF2016.html`, 2016. Accessed: Oct. 22, 2018.

[20] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5, 2017.

[21] 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.

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

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

[24] 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.

[25] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modeling & Visualization*, pages 225–226, 2013.

[26] Youngjin Yoon, Haegon Jeon, Donggeun Yoo, Joonyoung Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 57–65, 2015.

[27] 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.

[28] Jingyi Yu, Xu Hong, Jason Yang, and Yi Ma. Dgene: The light of science, the light of future. `http://www.plex-vr.com/product/model/`. Accessed: Oct. 22, 2018.

[29] Shuo Zhang, Hao Sheng, Da Yang, Jun Zhang, and Zhang Xiong. Micro-lens-based matching for scene recovery in lenslet cameras. *IEEE Transactions on Image Processing*, 27(3):1060–1075, 2018.

[30] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.