High-Order Residual Network for Light Field Super-Resolution

Nan Meng,[1*] **Xiaofei Wu,**[2] **Jianzhuang Liu,**[2] **Edmund Y. Lam**[1]
[1]Dept. of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong
[2]Huawei Noah's Ark Lab, China
{nanmeng, elam}@eee.hku.hk, {wuxiaofei2, liu.jianzhuang}@huawei.com

Abstract

Plenoptic cameras usually sacrifice the spatial resolution of their SAIs to acquire geometry information from different viewpoints. Several methods have been proposed to mitigate such spatio-angular trade-off, but seldom make use of the structural properties of the light field (LF) data efficiently. In this paper, we propose a novel high-order residual network to learn the geometric features hierarchically from the LF for reconstruction. An important component in the proposed network is the high-order residual block (HRB), which learns the local geometric features by considering the information from all input views. After fully obtaining the local features learned from each HRB, our model extracts the representative geometric features for spatio-angular upsampling through the global residual learning. Additionally, a refinement network is followed to further enhance the spatial details by minimizing a perceptual loss. Compared with previous work, our model is tailored to the rich structure inherent in the LF, and therefore can reduce the artifacts near non-Lambertian and occlusion regions. Experimental results show that our approach enables high-quality reconstruction even in challenging regions and outperforms state-of-the-art single image or LF reconstruction methods with both quantitative measurements and visual evaluation.

Compared to a 2D imaging system, the plenoptic camera not only captures the accumulated intensity of a light ray at each point in space, but also provides the directional radiance information. Together they form the light field (LF), which has shown advantages over 2D imagery in problems such as disparity estimation (Jeon et al. 2015; Sun et al. 2016) or 3D reconstruction (Heber, Yu, and Pock 2017) of a scene, generation of images for a novel viewpoint (Kalantari, Wang, and Ramamoorthi 2016; Meng et al. 2019b), and refocusing (Mitra and Veeraraghavan 2012).

Nevertheless, in practice, it can be difficult to achieve a dense sampling of the entire LF due to the limited resolution of the camera sensor. Acquisition of a densely sampled subaperture image (SAI) usually sacrifices the view point information, or vice versa (Wanner and Goldluecke 2014). As a result, the LF views exhibit a lower spatial resolution than images obtained by conventional cameras, and many applications such as depth estimation are constrained by low-resolution (LR) images, which increases the significance of the algorithms for light field super-resolution (LFSR).

Different from single image super-resolution (SISR), a LF is characterized by a structure that needs to be maintained when increasing the data resolution. Such structural information is implicitly encoded in the neighboring views, leading to non-integral shift between two corresponding pixels in the view images (Wu et al. 2017a). From this perspective, most depth-based approaches (Wanner and Goldluecke 2014; Mitra and Veeraraghavan 2012) generally depend on such geometric properties as priors to explicitly register the novel SAIs from other views. Their success depend on the accurate depth information, which however is challenging to acquire. Consequently, the disparity errors give rise to artifacts such as tearing and ghosting, especially in the occluded or non-Lambertian areas where depth information is not properly estimated.

Recently, deep learning has shown to be powerful in various computer vision applications (Meng et al. 2018; Yang, Chen, and Shao 2019), including LFSR (Wu et al. 2017b; Meng, Zeng, and Lam 2019; Meng et al. 2019a). The learning-based approaches relieve the dependency on explicit depth information, leading to the improvement of robustness at depth discontinuities. However, the intrinsic limitation of 2D (or 3D) convolution makes existing frameworks difficult to handle the high-dimensional structure in a LF, and therefore most learning-based algorithms simplify the reconstruction to consider only the spatio-angular relations in epipolar plane images (EPIs) (Wu et al. 2017b), or the angular correlations among adjacent views (Yoon et al. 2017; Zhang, Lin, and Sheng 2019). Given that the geometric information is encoded in a complex way within the LF, such simplifications result in performance degradation.

To remedy the problems of existing learning-based approaches for LFSR, we propose to establish a framework tailored to the LF structural information. Such an approach enables the network to learn representations by fully exploiting the LF information from all adjacent views. The main contributions of our model are threefold: 1) We propose a novel high-order structure, named high-order residual block

[*]This work was done during an internship at Huawei.

(HRB) to learn the features by fully considering the information from all SAIs of a LF. Such features extracted from HRB preserve high angular coherence. 2) By stacking a set of the HRBs, the proposed network is able to extract diverse spatial features endowed with scene geometry information. In addition, the network also propagates the geometric information encoded in the learned features to achieve high reconstruction quality with the LF structural property. 3) Experimental results demonstrate that our model not only outperforms the state-of-the-art reconstruction methods on quantitative measurements but also generates spatial details and novel views with better fidelity.

Related Work

Spatial super-resolution

A number of super-resolution algorithms (Bishop and Favaro 2012; Lim et al. 2009; Vagharshakyan, Bregovic, and Gotchev 2018) have been developed specifically for LF data. For example, in (Wanner and Goldluecke 2014), a variational framework is applied to super-resolve a novel view based on a continuous depth map calculated from the epipolar plane images (EPIs). Mitra and Veeraraghavan (Mitra and Veeraraghavan 2012) proposed a patch-based approach, which is also based on the estimated depth information. These methods usually require an accurate estimate of the disparity information, which can be challenging on LR images. Learning-based approaches mitigate the dependency on geometric disparity, and therefore are more robust in regions where the depth information is difficult to estimate correctly. In (Yoon et al. 2017), LFCNN cascades two CNNs to enhance the target views and generate novel perspectives based on the super-resolved views. However, the stepwise processing does not make use of the entire structural information of the LF, and therefore limits the potential of the model. Recently, (Wang et al.) employed a bidirectional recurrent CNN framework to model the spatial correlations horizontally and vertically. Likewise, (Farrugia, Galea, and Guillemot 2017) adopted an example-based spatial SR algorithm on the patch-volumes across the SAIs. Both approaches consider the LF as an image sequence, and therefore lose one angular dimension information. (Zhang, Lin, and Sheng 2019), however, utilized the relations of SAIs in 4 different directions to super-resolve the center target image. By considering the angular information from multiple directions, their model has superior performance over other previous methods.

Angular super-resolution

Angular super-resolution for LF is also known as view synthesis. Many techniques (Kalantari, Wang, and Ramamoorthi 2016; Meng et al. 2019b; Wanner and Goldluecke 2014) take advantage of the disparity map to warp the existing SAIs to novel views. For instance, (Pearson, Brookes, and Dragotti 2013) introduced a layer-based synthesis method to render arbitrary views by using probabilistic interpolation and calculating the depth layer information. (Zhang, Liu, and Dai 2015) adopted a phase-based method, which integrates the disparity into a phase term of a reference image to warp the input view to any close novel view.

Similar to spatial super-resolution, depth-based techniques are inadequate in the occluded and textureless regions, which prompts researchers to explore algorithms based on CNNs. (Flynn et al. 2016) are among the first to apply deep learning to view synthesis from a set of images with wide baselines. Meanwhile, (Kalantari, Wang, and Ramamoorthi 2016) exploited two sequential CNNs to estimate depth and color information, and subsequently wrapped them to generate the novel view. The dependency on disparity restricts the model performance and easily results in ghosting effects near occluded regions. In (Wu et al. 2017b), the author proposed to use a blur-deblur scheme to address the asymmetry problem caused by sparse angular sampling. However, this EPI-based model only utilizes horizontal or vertical angular correlations of a low-resolution LF, which severely restricts the accessible information of the model. Recently, (Wing Fung Yeung et al. 2018) applied the alternating convolution to learn the spatio-angular clues for view synthesis and achieves more accurate results.

Compared with the aforementioned approaches, we explore a deeper residual structure for both spatial and angular SR of the LF. The proposed network can harness the high-dimensional LF data efficiently to extract geometric features, which contribute to the high reconstruction accuracy.

Method

Problem formulation

Following (Levoy and Hanrahan 1996; Gortler et al. 1996), a light ray is defined by the intersection points of an angular plane (s, t) and a spatial plane (x, y). We consider the LFSR as the recovery of the HR LF $I^H(x, y, s, t) \in R^{\gamma_s X \times \gamma_s Y \times \gamma_a S \times \gamma_a T}$ from the input LR LF $I^L(x, y, s, t) \in R^{X \times Y \times S \times T}$ by two spatial and angular SR factors γ_s and γ_a, respectively. The learning-based SR process can be described as

$$I^S(x, y, s, t) = g(I^L(x, y, s, t); \Theta), \tag{1}$$

where I^S stands for the super-resolved LF, $g(\cdot)$ represents the mapping from LR to super-resolved LF, and Θ denotes the parameters of the model.

Architecture overview

The intrinsic limitation of 2D and 3D convolution makes existing schemes unable to fully exploit highly-correlated 4D LF data. As a result, most existing methods consider only partial spatio-angular relations (e.g., EPI) (Wu et al. 2017b; 2018), or angular correlations (e.g., SAI) (Kalantari, Wang, and Ramamoorthi 2016; Wang et al. ; Zhang, Lin, and Sheng 2019) which underuse the potential of LF. To resolve the problem, we employ a high-order convolution (HConv) that encapsulates the information from all coordinates by convolving a 4D kernel with the inputs. For any hidden layer $\mathbf{H}^{(k)}$ $(k \in \{1, 2, \ldots, K\})$, the operation of the HConv (together with the following activation layer) is implemented as $\mathbf{H}^{(k)} = \delta(\mathbf{W}^k \star \mathbf{H}^{(k-1)})$. $\mathbf{W}^{(k)}$ denotes the weights of the

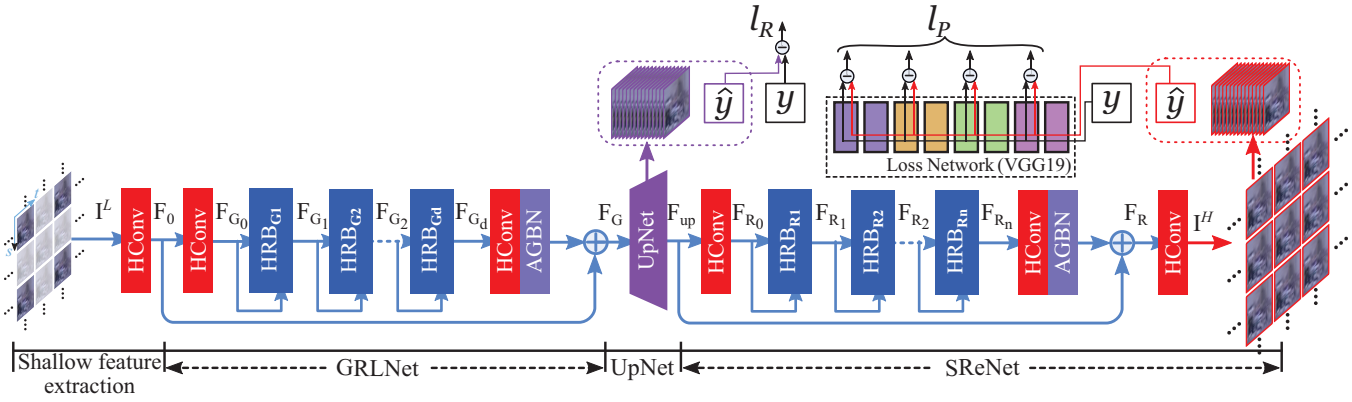


Figure 1: The architecture of our proposed hierarchical high-order network.

k^{th} layer with size $s_1 \times s_2 \times a_1 \times a_2 \times c$, where c is the channel number of the filter bank, $s_1 \times s_2$ is the spatial filter size and $a_1 \times a_2$ is the angular filter size. $\mathbf{H}^{(0)}$ stands for the input I^L, and the activation function $\delta(\cdot)$ is the leaky rectified linear unit (LReLU) with slope $\alpha = 0.2$. The notation $\star$ is the convolution between an input feature map and a filter. Furthermore, to utilize spatial information hierarchically from all SAIs, we design the high-order residual block (HRB) to effectively extract the *geometric features* from a LF. As is shown in Fig. 1, the proposed network mainly consists of four parts: 1) shallow feature extraction, 2) geometric representation learning network (GRLNet), 3) upsampling network (UpNet), and 4) spatial refinement network (SReNet). Specifically, we use a HConv layer to extract shallow features F_0 from the LR input:

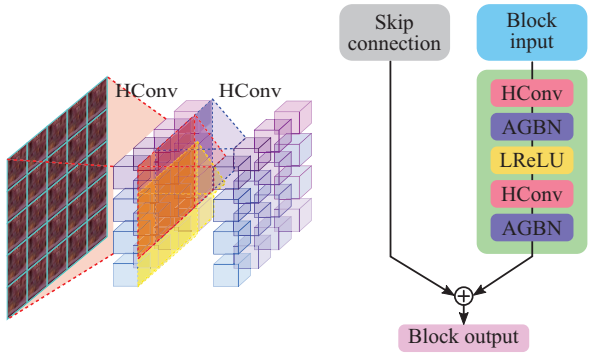
$$F_0 = H_{\text{HC}}(I^L), \tag{2}$$

where $H_{\text{HC}}(\cdot)$ denotes the HConv operation. Subsequently, in the GRLNet, the representations are learned in a hierarchical manner by a set of HRBs. Assuming there are d HRBs, the feature maps F_{G_d} of the d^{th} HRB can be expressed as:

$$\begin{aligned} F_{\text{G}_\text{d}} &= H^{\text{d}}_{\text{HRB}}(F_{\text{G}_{\text{d}-1}}) \\ &= H^{\text{d}}_{\text{HRB}}(H^{\text{d}-1}_{\text{HRB}}(F_{\text{G}_{\text{d}-2}})) \\ &= H^{\text{d}}_{\text{HRB}} \circ H^{\text{d}-1}_{\text{HRB}} \circ \cdots \circ H^{1}_{\text{HRB}}(F_{\text{G}_0}), \end{aligned} \tag{3}$$

where $H^{\text{d}}_{\text{HRB}}(\cdot)$ denotes the operation of the d^{th} HRB, and the symbol $\circ$ stands for function composition. The mapping $H^{\text{i}}_{\text{HRB}}(\cdot), (i = 1, 2, \ldots, d)$ can be a composite function of operations, including 2 HConvs to fully utilize all the view information within the block (Fig. 2) to obtain the local geometric feature F_{G_i}. By cascading multiple HRBs, the geometric features are learned in an hierarchical manner during the training of GRLNet, and therefore more representative features with diverse spatial representations are obtained. We then apply the global residual learning to combine the hierarchically learned geometric features F_{G_d} and the shallow features F_0 before conducting upsampling by

$$F_{\text{G}} = H_{\text{AGBN}} \circ H_{\text{HC}}(F_{\text{G}_\text{d}}) + F_0, \tag{4}$$



(a) The angular receptive field of HRB (b) The structure of HRB

Figure 2: The high-order residual block (HRB) architecture, where $\oplus$ denotes the element-wise addition.

where H_{AGBN} is an operation of batch normalization defined later. The following UpNet then upsamples the obtained feature maps F_{G} from the LR space to the HR space:

$$F_{\text{up}} = H_{\text{up}}(F_{\text{G}}), \tag{5}$$

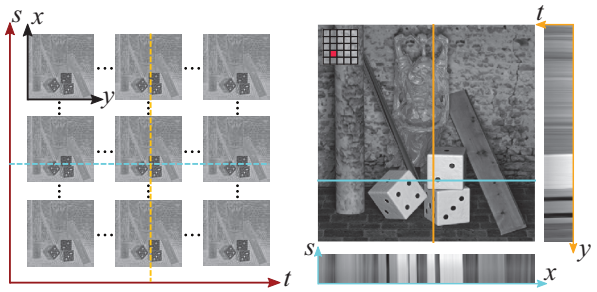
where H_{up} is used to describe the upsampling operation on the LR features. In the experiments, however, directly reconstructing the HR LF based on the fused features is hard, and the results always lack high-frequency spatial details. Therefore, we employ a refinement network (SReNet) supervised by a perceptual loss to recover the spatial details in the HR space:

$$F_{\text{R}} = H_{\text{AGBN}} \circ H_{\text{HC}}(F_{\text{R}_\text{n}}) + F_{\text{up}}, \tag{6}$$

where $n \leq d$ and F_{R} denotes the fused refined feature, which is further used for the reconstruction of the final super-resolved LF:

$$I^H = H_{\text{HC}}(F_{\text{R}}). \tag{7}$$

In the following sections, we will illustrate the components of the proposed high-order network in details, and demonstrate the properties of learned geometric features.



(a) 2D feature slices (b) Feature EPIs

Figure 3: Visualization of the geometric features. (a) The collection of 2D slices through the learned feature maps. (b) The EPI located at the corresponding lines.

High-order residual block

As a basic building block in our network, the HRB's structure is presented in Fig. 2. According to the Fig. 2a, each HRB contains two HConv layers with the 3×3 angular receptive field which makes the block possible to fully utilize the information from all SAIs of input features. In addition, to ease the training of the proposed high-order network, we apply the normalization operation to the outputs of the HConv layer (Ioffe and Christian 2015). Given that the inputs preserve high coherence among the views, the normalization should not be counted in an aperture-wise manner to avoid that the whitening decorrelates the coherence. We consequently implement the normalization over a group of SAIs in every channel of the feature maps and therefore propose an aperture group batch normalization (AGBN).

Let the outputs of a particular hidden HConv layer be $\mathbf{H} = \{\mathcal{H}_m^n(s,t)\}$, where s and t stand for the two indices of angular dimensions as is defined in the problem formulation. The superscript $n \in [1, N]$ denotes the number of channels, and for each feature SAI contains M values ($m = 1, 2, \ldots, M$). Therefore, the AGBN transform is implemented as in Algorithm 1.

Algorithm 1: Aperture group batch normalization

Input: Features from HConv layer: $\mathbf{H} = \{\mathcal{H}^n(s,t)\}$; Parameters γ and β

Output: The output features: $\hat{\mathbf{H}} = \{\hat{\mathcal{H}}^n(s,t)\}$

for $n = 1, 2, \ldots, N$ **do**

$$\mu_n \leftarrow \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{S\cdot T}\sum_{s=1}^{S}\sum_{t=1}^{T}\mathcal{H}_m^n(s,t)\right) = \frac{1}{MST}\sum_{m=1}^{M}\sum_{s=1}^{S}\sum_{t=1}^{T}\mathcal{H}_m^n(s,t));$$

$$\sigma_n \leftarrow \frac{1}{MST}\sum_{m=1}^{M}\sum_{s=1}^{S}\sum_{t=1}^{T}(\mathcal{H}_m^n(s,t)-\mu_n)^2;$$

$$\hat{\mathcal{H}}^n(s,t) \leftarrow \gamma \cdot \frac{\mathcal{H}^n(s,t)-\mu_n}{\sqrt{\sigma_n^2+\epsilon}} + \beta$$

end

By stacking the layers as is shown in Fig. 2b, the HRB is able to extract features that preserve geometrical properties by considering all SAIs. The learned geometrical features not only contain spatial structures (such as textures or edges) but also record the relations between adjacent feature views. Fig. 3 exhibits an example of the geometric features learned by the HRB. To illustrate such high-dimensional features, we display a grid of 2D slices through the 4D features in Fig. 3a, and the EPIs located at the corresponding lines in a certain feature slice in Fig. 3b. The feature EPIs very much resemble the LF EPIs, reflecting that the HRB has the capacity to extract features preserving high coherence.

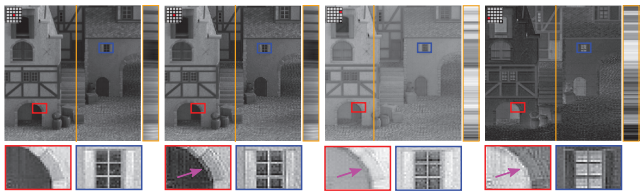


Figure 4: Illustration of the geometric features extracted from different HRBs in the GRLNet.

Geometric representation learning network

The GRLNet is composed of a set of cascaded HRBs. Such structure enables the network to learn multiple spatial representations endowed with geometry information. Compared with the features extracted from traditional CNN-based model, the learned geometric features are different in two aspects: 1) the high coherence among SAIs in angular dimension; 2) the *smooth* effects near object borders in spatial dimension. The former has been discussed in Fig. 3b, where we show the *EPI property* of features. The latter can be illustrated according to the features from different HRBs. In Fig. 4, we visualize the spatial appearance of the geometric features extracted from the 3^{rd}, 5^{th}, and 8^{th} HRBs. The red boxes zoom in at the features of object border, while the blue boxes zoom in at the texture features. Compared with the reconstruction result, the edges of the object border in the feature space (red boxes) are not as sharp. Such effects are caused by the rapid changes in the parallax of the object border, and therefore indicate the scene geometric information. In addition to the diverse spatial representations, the feature angular coherence is also maintained (e.g., the EPI in the yellow boxes). Consequently, the GRLNet is able to learn more representative spatial features hierarchically through a set of HRBs and simultaneously propagates the geometric information.

Upsampling network

The upsampling network is applied to increase the spatio-angular resolution using the extracted hierarchical geometric features in the LR space. We design the network to fit to the properties of the LF geometric features. As illustrated in Fig. 5, we assume a single LR feature map with a single channel as input. The feature map has dimension $X \times Y \times S \times T$, where $X = Y = 3$ and $S = T = 3$. The spatial and angular upsampling factors are $\gamma_s = \gamma_a = 2$ (strictly speaking, the angular dimension is increased from 3×3 to 5×5). The first step expands the feature channel by a factor of γ_s^2 using the HConv operation. Then, given

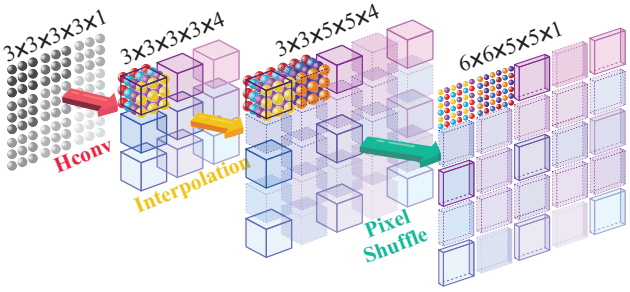


Figure 5: Illustration of UpNet for spatio-angular resolution enhancement. The red arrow stands for HConv opeartion, the yellow one denotes angular linear interpolation, and the green one denotes channel-to-space pixel shuffle operation.

the EPI property of the geometric features, we apply a linear interpolation to the angular dimension of the feature maps to upscale the resolution by a factor of γ_a. Finally, the channel-to-space shuffle operation is applied to increase the spatial resolution by a fator of γ_s.

The UpNet upsamples the learned geometric features to the HR space. Such features are used to reconstruct the primary super-resolved LF directly to get the per-pixel reconstruction loss for training, and are also passed to the SReNet to further recover the high-frequency details.

Spatial refinement network

The SReNet aims at restoring the realistic spatial details of the previous super-resolved output. Given that the GRL-Net is trained using a pixel-wise loss, it tends to generate smooth results with poor fidelity (Ledig et al. 2017; Gupta et al. 2011). The SReNet in contrast learns the geometric features directly in the HR space and is supervised by a novel perceptual loss defined on SAIs to make the reconstruction sharper. In the experiments, we will discuss the effects of the SAI-wise perceptual loss for recovering spatial details.

Loss function

We propose a two-stage loss function (see Fig. 1) to encourage the proposed network to learn the geometric features and reconstruct high-quality spatial details. In general, the loss function is a linear combination of two terms:

$$\ell = \alpha \cdot \ell_R + \beta \cdot \ell_P. \tag{8}$$

The *reconstruction loss* ℓ_R models the pixel-wise difference between the super-resolved LF I^S and the ground truth I^H:

$$\ell_R = \sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{s=1}^{S}\sum_{t=1}^{T}(I^H(x,y,s,t) - I^S(x,y,s,t))^2. \tag{9}$$

The *perceptual loss* ℓ_P measures the quality of the spatial reconstruction. Inspired by (Johnson, Alahi, and Fei-Fei 2016), we define the loss function acquired from a VGG network to describe the aperture-wise differenes between high-level features ϕ,

$$\ell_P = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T}\left(\phi(I^H_{s,t}) - \phi\left(g(I^L_{s,t};\Theta)\right)\right)^2, \tag{10}$$

where $I^L_{s,t} = I^L(\cdot,\cdot,s,t)$ and $I^H_{s,t} = I^H(\cdot,\cdot,s,t)$ denote the LR input and the ground truth with angular coordinate (s,t), respectively.

Experiments

Data and experiment settings

In the experiments, we randomly select 100 scenes from the Lytro Archive (Stanford) (excluding "Occlusions" and "Reflective") and the entire Fraunhofer densely-sampled high-resolution (Ziegler et al. 2017) datasets for training. The former consists of 353 real-world scenes captured using a Lytro Illum camera with a small baseline, and in addition, we exclude the corner samples and only select the center 9×9 views in the experiments. The latter contains 9 real-world scenes that are densely sampled by a high-resolution camera with a larger baseline. The experimental results show that our trained network can be generalized to various synthetic and real-world scenes, as well as some microscopy light fields. This indicates that the learned geometric features are generic for various situations.

During training, the system each time receives a 4D patch of LF, which is spatially cropped to 96×96 as input. For spatial SR, the downsampling is based on the classical model (Farrugia and Guillemot 2018)

$$I^L = \downarrow_{\gamma_s} G * I^H + \eta, \tag{11}$$

where η is Gaussian noise with zero mean and unit standard deviation, $\downarrow_{\gamma_s}$ denotes the nearest neighbor downsampling operator applied to each view, γ_s is the magnification factor, and G stands for a Gaussian blurring kernel with a window size of 7×7 and standard deviation of 1.2 pixels. The network is trained using the Stochastic Gradient Descent solver with the initial learning rate of 10^{-5}, which is decreased by a factor of 0.1 for every 10 epochs. The entire implementation is available at *https://github.com/monaen/LightFieldReconstruction*.

Loss evaluation

To exam the effectiveness of different loss components, we adjust the proposed network to obtain multiple variants which are further trained using different losses. The parameters of different variants are kept constant to control the model representational capacity. In Table 2, we evaluate the performance of the variants with 8 HRBs in total. By combining the reconstruction and perceptual losses, the model can achieve comprehensively better quantitative results and reconstruct the LF with good visual fidelity (refer to the supplementary materials).

Spatial super-resolution evaluation

For evaluation in terms of spatial resolution, we compare against several top-performing algorithms designed for LFSR, including LFCNN (Yoon et al. 2017), BM PCA+RR (Farrugia, Galea, and Guillemot 2017), LFNet (Wang et al.), Zhang et al. (Zhang, Lin, and Sheng 2019) and some state-of-the-art methods for SISR, like MSLapSRN (Lai et al. 2018) and RDN (Zhang et al. 2018). For fair comparison, all the methods are retrained using

Table 1: Quantitative evaluation of state-of-the-art methods for spatial and angular LFSR. We report the average PSNR and SSIM over all sub-aperture images for Spatial $2\times$, $3\times$, $4\times$ and Angular $3\times$ ($3 \times 3 \rightarrow 9 \times 9$). The **bold** values indicate the best performance.

Algorithm	Scale	PSNR(dB)					SSIM				
		Occlusions (20)	Reflective (20)	HCI new (8)	HCI old	EPFL (21)	Occlusions (20)	Reflective (20)	HCI new (8)	HCI old	EPFL (21)
Bicubic	S×2	28.52	31.19	29.69	26.97	28.66	0.808	0.863	0.806	0.769	0.849
LFCNN		28.86	31.42	31.24	28.41	29.69	0.834	0.885	0.851	0.826	0.864
BM PCA+RR		30.45	33.07	32.62	31.20	32.68	0.878	0.896	0.879	0.892	0.906
LFNet		30.37	33.85	32.81	29.56	32.66	0.881	0.912	0.898	0.884	0.892
Zhang et al.		34.28	36.31	34.14	32.17	33.70	0.931	0.946	0.910	0.921	0.939
MSLapSRN		30.85	32.43	33.13	29.51	32.27	0.879	0.909	0.894	0.852	0.901
RDN		31.46	33.86	33.25	31.13	32.41	0.893	0.916	0.893	0.894	0.912
Ours		**35.13**	**37.41**	**34.85**	**33.29**	**36.24**	**0.937**	**0.951**	**0.916**	**0.934**	**0.949**
Bicubic	S×3	26.95	29.50	28.94	25.39	27.31	0.746	0.819	0.776	0.703	0.812
LFCNN		27.25	29.78	29.75	25.79	27.51	0.758	0.826	0.803	0.733	0.827
BM PCA+RR		27.96	30.05	30.24	26.78	29.71	0.817	0.835	0.834	0.766	0.850
LFNet		28.01	30.34	29.81	26.76	29.69	0.816	0.847	0.822	0.764	0.848
Zhang et al.		29.85	32.30	28.33	28.24	30.16	0.840	0.881	0.738	0.807	0.871
MSLapSRN		29.22	32.03	30.94	27.80	30.00	0.818	0.875	0.828	0.789	0.867
RDN		29.14	30.82	29.54	26.89	29.65	0.796	0.846	0.788	0.755	0.834
Ours		**31.18**	**33.55**	**32.00**	**29.40**	**32.84**	**0.871**	**0.903**	**0.867**	**0.849**	**0.906**
Bicubic	S×4	24.98	27.54	25.92	23.95	25.94	0.663	0.771	0.688	0.630	0.767
LFCNN		25.04	28.14	28.28	25.65	26.97	0.686	0.798	0.768	0.688	0.792
BM PCA+RR		26.28	28.73	28.90	25.85	27.51	0.710	0.796	0.772	0.703	0.785
LFNet		25.94	28.81	29.36	25.40	26.10	0.709	0.808	0.762	0.706	0.775
Zhang et al.		27.36	29.69	28.74	25.83	28.02	0.744	0.810	0.743	0.694	0.798
MSLapSRN		27.41	30.28	29.55	26.27	28.78	0.755	0.835	0.782	0.723	0.821
RDN		26.97	29.64	29.63	26.66	28.58	0.724	0.817	0.792	0.730	0.804
Ours		**29.04**	**30.82**	**30.78**	**28.54**	**29.63**	**0.815**	**0.859**	**0.837**	**0.821**	**0.859**
Kalantari et al.	A×3	34.70	37.24	35.53	32.59	35.19	0.927	0.958	0.916	0.906	0.959
Wu et al.		35.64	40.03	35.64	33.38	37.05	0.928	0.963	**0.918**	0.905	0.960
Ours (4HRBs)		**35.96**	**40.16**	**35.79**	**33.75**	**38.28**	**0.930**	**0.964**	0.904	**0.913**	**0.961**

Table 2: Ablation study of different components in the proposed model. "G8" denotes 8 HRBs in GRLNet, "S8" denotes 8HRBs in SReNet, and "G5S3" stands for 5 HRBs in GRLNet and 3 HRBs in SReNet.

Model	GRLNet	SReNet	Loss	HCI new	Occlusions (20)
G8	√	×	ℓ_R	31.35	32.70
S8	×	√	ℓ_P	31.34	32.65
G5S3	√	√	ℓ_R+ℓ_P	31.34	32.76

the same datasets and downsampling method described in Eq. 11. Table 1 shows the quantitative comparisons for $\times 2$, $\times 3$, and $\times 4$ SR on five public LF dataset. The real-world datasets consist of 20 scenes from "Occlusions" and 20 scenes from "Reflective" in Stanford Lytro Archive (Stanford), and 21 scenes from EPFL (Rerabek and Ebrahimi 2016), while the synthetic datasets are selected from the HCI dataset (Honauer et al. 2016; Wanner, Meister, and Goldluecke). We carefully fine-tune and retrain each algorithm to fit the classic downsampling method described in Eq. 11 using their publicly available code to reach their best performance. The results are measured in terms of the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) over the center 9×9 views of each evaluation scene, and we report the average value on each test dataset.

Fig. 6 compares the visual reconstruction results for $4\times$ spatial SR. For LFSR algorithms, LFCNN receives pairs of SAIs as input without modeling their correlations. Therefore, it underuses the angular information and tends to generate over-smoothed results with blurry details. Likewise, the two SISR methods (MSLapSRN and RDN) do not model the correlations either, which leads to the blurring and twisting in their EPIs. BM PCA+RR and LFNet simplify the problem by considering only one dimension of angular correlations. Such strategy restricts the performance of their methods in terms of both visual results (Fig. 6) and quantitative measurements (Table 1). In contrast, a recent approach proposed by Zhang et al. (Zhang, Lin, and Sheng 2019) exploits the angular information from 4 directions for LF spatial SR. By integrating more directional information, the algorithm shows better quantitative results on spatial $2\times$ and $3\times$ tasks. Nevertheless, their approach fuses the angular information by roughly concatenating SAIs from 4 directions in the channel dimension. Given that the differences

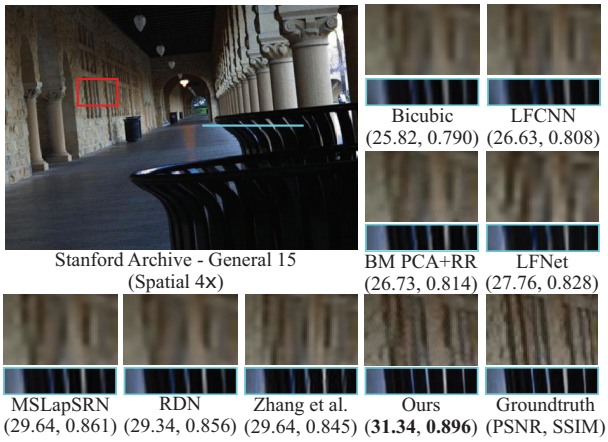


Figure 6: Visual comparison for $4\times$ spatial SR on the real-world scene General 15 from Stanford Archive.

between adjacent views decrease rapidly in the LR LF, its performance on $4\times$ spatial SR is badly affected. Compared with these methods, our model exploits the entire angular information from all directions. In addition, such angular information is further fused in the learned geometric features during the network training. In this way, all of the structural information is utilized for the final reconstruction allowing our model to achieve superior performance in terms of both visual fidelity (e.g., the texture on the wall in Fig. 6) and quantitative measurements. More visual results on both real-world and synthetic scenes are presented in our supplementary materials to illustrate the model generalization ability.

Angular super-resolution evaluation

For angular super-resolution, all the models are trained only using the 100 Lytro Archive scenes. We carry out comparisons with three state-of-the-art CNN based methods, namely, Kalantari et al. (Kalantari, Wang, and Ramamoorthi 2016), Wu et al. (Wu et al. 2017b) and Yeung et al. (Wing Fung Yeung et al. 2018). For all the angular SR task, we evaluate a simplified version of the proposed model with 3 HRBs in GRLNet and 1 HRB in SReNet (8 HConvs in total). Even with only 4 HRBs, our model is able to defeat the other three methods. For the $3 \times 3 \rightarrow 9 \times 9$ synthesis task, the quantitative comparisons on average PSNR and SSIM are presented in the last part of Table 1. Our model defeats Kalantari et al. and Wu et al. on most real-world and synthetic LF scenes. Fig. 7 compares the visual results. The depth-dependent method Kalantari et al. tends to produce ghosting artifacts near boundaries of objects. Wu et al. only uses the EPI information and therefore leads to a loss of spatial details (e.g., Neurons $20\times$ nerve fibre is absent in their reconstructed LF). To demonstrate the effectiveness of the hierarchical HRBs structures, we further compare the performance against Yeung et al.'s model with 16 4D alternating convolutions (16L). According to the Table 3 and Fig.7, our model achieves higher quantitative values and synthesizes more realistic novel views (also refer to our supplementary video).

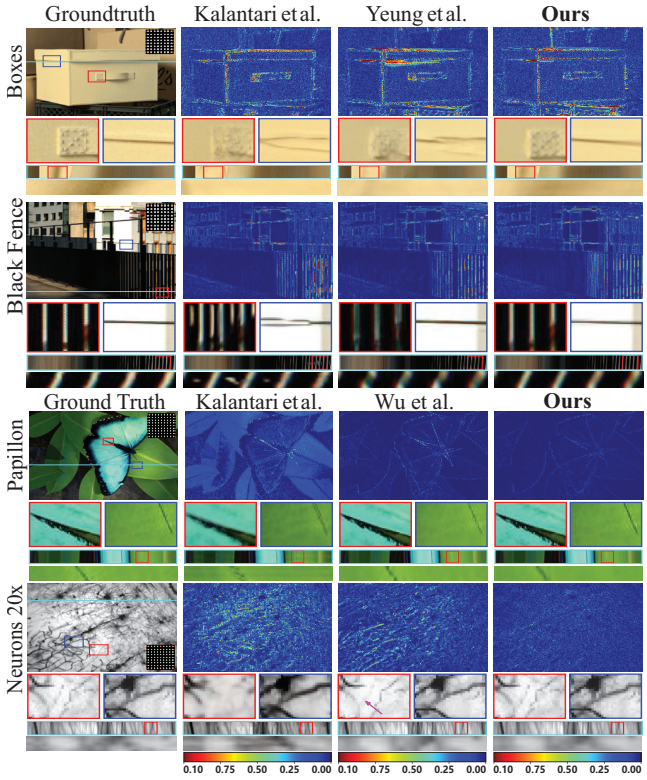


Figure 7: Visual comparison of our model with Kalantari et al. and Yeung et al. for the $2 \times 2 \rightarrow 8 \times 8$ task, and with Wu et al. for the $3 \times 3 \rightarrow 9 \times 9$ task.

Algorithm	Occlusions (20)	Reflective (20)	EPFL (21)	HCI new
Kalantari et al.	32.68	35.98	33.60	33.19
Yeung et al.(16L)	33.19	36.82	35.09	33.39
Ours (4HRBs)	**33.22**	**36.93**	**35.30**	**34.04**

Table 3: Quantitative evaluation of state-of-the-art view synthesis algorithms. We report the average PSNR for the task $2 \times 2 \rightarrow 8 \times 8$.

Conclusion

In this paper, we design a hierarchical high-order framework for LF spatial and angular SR. To fully exploit the structural information of the LF, HRB has been proposed. By cascading a set of HRBs, our model is able to extract representative features encoded with geometric information. Such features contribute a lot to the final reconstruction results. In addition, the combination of the pixel-wise loss and the perceptual loss further allows our model to generate more realistic spatial images. The experiments show that our proposed model outperforms the state-of-the-art SR methods in terms of both quantitative measurements and visual fidelity.

Acknowledgments

This work is supported in part by the Research Grants Council of Hong Kong (GRF 17203217, 17201818, 17200019)

and the University of Hong Kong (104005009, 104005438).

References

Bishop, T. E., and Favaro, P. 2012. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE TPAMI*.

Farrugia, R. A., and Guillemot, C. 2018. Light field super-resolution using a low-rank prior and deep convolutional neural networks. *arXiv preprint*.

Farrugia, R. A.; Galea, C.; and Guillemot, C. 2017. Super resolution of light field images using linear subspace projection of patch-volumes. *IEEE JSTSP*.

Flynn, J.; Neulander, I.; Philbin, J.; and Snavely, N. 2016. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*.

Gortler, S. J.; Grzeszczuk, R.; Szeliski, R.; and Cohen, M. F. 1996. The lumigraph. In *Siggraph*.

Gupta, P.; Srivastava, P.; Bhardwaj, S.; and Bhateja, V. 2011. A modified PSNR metric based on HVS for quality assessment of color images. In *ICCIA*.

Heber, S.; Yu, W.; and Pock, T. 2017. Neural EPI-volume networks for shape from light field. In *ICCV*.

Honauer, K.; Johannsen, O.; Kondermann, D.; and Goldluecke, B. 2016. A dataset and evaluation methodology for depth estimation on 4D light fields. In *ACCV*.

Ioffe, S., and Christian. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint*.

Jeon, H.-G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.-W.; and So Kweon, I. 2015. Accurate depth map estimation from a lenslet light field camera. In *CVPR*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Kalantari, N. K.; Wang, T.-C.; and Ramamoorthi, R. 2016. Learning-based view synthesis for light field cameras. *ACM TOG*.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2018. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE TPAMI*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*.

Levoy, M., and Hanrahan, P. 1996. Light field rendering. In *ACM CCGIT*.

Lim, J.; Ok, H.; Park, B.; Kang, J.; and Lee, S. 2009. Improving the spatail resolution based on 4D light field data. In *ICIP*.

Meng, N.; Lam, E.; Tsia, K. K. M.; and So, H. K.-H. 2018. Large-scale multi-class image-based cell classification with deep learning. *IEEE JBHI*.

Meng, N.; So, H. K.-H.; Sun, X.; and Lam, E. 2019a. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE TPAMI*.

Meng, N.; Sun, X.; So, H. K.-H.; and Lam, E. Y. 2019b. Computational light field generation using deep nonparametric bayesian learning. *IEEE Access*.

Meng, N.; Zeng, T.; and Lam, E. Y. 2019. Spatial and angular reconstruction of light field based on deep generative networks. In *ICIP*.

Mitra, K., and Veeraraghavan, A. 2012. Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior. In *CVPRW*.

Pearson, J.; Brookes, M.; and Dragotti, P. L. 2013. Plenoptic layer-based modeling for image based rendering. *IEEE TIP*.

Rerabek, M., and Ebrahimi, T. 2016. New light field image dataset. In *QoMEX*.

Sun, X.; Xu, Z.; Meng, N.; Lam, E. Y.; and So, H. K.-H. 2016. Data-driven light field depth estimation using deep convolutional neural networks. In *IJCNN*.

Vagharshakyan, S.; Bregovic, R.; and Gotchev, A. 2018. Light field reconstruction using shearlet transform. *IEEE TPAMI*.

Wang, Y.; Liu, F.; Zhang, K.; Hou, G.; Sun, Z.; and Tan, T. LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE TIP*.

Wanner, S., and Goldluecke, B. 2014. Variational light field analysis for disparity estimation and super-resolution. *IEEE TPAMI*.

Wanner, S.; Meister, S.; and Goldluecke, B. Datasets and benchmarks for densely sampled 4D light fields. In *VMV*.

Wing Fung Yeung, H.; Hou, J.; Chen, J.; Ying Chung, Y.; and Chen, X. 2018. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *ECCV*.

Wu, G.; Masia, B.; Jarabo, A.; Zhang, Y.; Wang, L.; Dai, Q.; Chai, T.; and Liu, Y. 2017a. Light field image processing: An overview. *IEEE JSTSP*.

Wu, G.; Zhao, M.; Wang, L.; Dai, Q.; Chai, T.; and Liu, Y. 2017b. Light field reconstruction using deep convolutional network on EPI. In *CVPR*.

Wu, G.; Liu, Y.; Fang, L.; Dai, Q.; and Chai, T. 2018. Light field reconstruction using convolutional network on EPI and extended applications. *IEEE TPAMI*.

Yang, Y.; Chen, H.; and Shao, J. 2019. Triplet enhanced autoencoder: model-free discriminative network embedding. In *IJCAI*.

Yoon, Y.; Jeon, H.-G.; Yoo, D.; Lee, J.-Y.; and Kweon, I. S. 2017. Light-field image super-resolution using convolutional neural network. *IEEE SPL*.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018. Residual dense network for image super-resolution. In *CVPR*.

Zhang, S.; Lin, Y.; and Sheng, H. 2019. Residual networks for light field image super-resolution. In *CVPR*.

Zhang, Z.; Liu, Y.; and Dai, Q. 2015. Light field from micro-baseline image pair. *CVPR*.

Ziegler, M.; op het Veld, R.; Keinert, J.; and Zilly, F. 2017. Acquisition system for dense lightfield of large scenes. In *CTVCTD*.