LFNet: A Novel Bidirectional Recurrent Convolutional Neural Network for Light-Field Image Super-Resolution

Yunlong Wang, *Student Member, IEEE*, Fei Liu, *Member, IEEE*, Kunbo Zhang, Guangqi Hou, Zhenan Sun, *Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

***Abstract*— The low spatial resolution of light-field image poses significant difficulties in exploiting its advantage. To mitigate the dependency of accurate depth or disparity information as priors for light-field image super-resolution, we propose an implicitly multi-scale fusion scheme to accumulate contextual information from multiple scales for super-resolution reconstruction. The implicitly multi-scale fusion scheme is then incorporated into bidirectional recurrent convolutional neural network, which aims to iteratively model spatial relations between horizontally or vertically adjacent sub-aperture images of light-field data. Within the network, the recurrent convolutions are modified to be more effective and flexible in modeling the spatial correlations between neighboring views. A horizontal sub-network and a vertical sub-network of the same network structure are ensembled for final outputs via stacked generalization. Experimental results on synthetic and real-world data sets demonstrate that the proposed method outperforms other state-of-the-art methods by a large margin in peak signal-to-noise ratio and gray-scale structural similarity indexes, which also achieves superior quality for human visual systems. Furthermore, the proposed method can enhance the performance of light field applications such as depth estimation.**

***Index Terms*— Implicitly multi-scale fusion, bidirectional recurrent convolutional neural network, light-field, super-resolution.**

I. Introduction

LIGHT Field (LF) originates from the concept of plenoptic function [1]–[6] and has recently come into the spotlight, especially with the emergence of commercial LF cameras [7], [8] and the dedication in the Virtual Reality (VR) field [9]. Compared to conventional cameras, a LF camera is capable of capturing both intensity values and directions of rays from real-world scenes. With the additional optical components like microlens inserted between the main lens and the image sensor, the rays crossing the main aperture deposit on pixels according to their spatial positions and propagation directions. Therefore it can capture a scene from multiple views in a single photographic exposure. While LF cameras provide high angular sampling for many computer vision applications [10]–[13], they are coherently spatially-undersampled [14]–[18] due to the trade-off between spatial and angular resolution. The low spatial resolution of LF image poses significant difficulties in exploiting its advantage.

Considering the narrow baseline between two neighbouring views of LF image, the parallax is generally around a few pixels according to the scene. Hence, LF image exhibits high correlations among the sub-aperture images. From this perspective, most of the preceding frameworks for Light-Field image Super-Resolution (LFSR) [15]–[20] generally depend on prior geometry information as sophisticated image priors to explicitly warp or register the sub-aperture images from slightly shifted views. However, the defect lies in that they basically require accurate geometric information of the scene as priors. Though numerous frameworks have been presented for depth estimation from LF image [21]–[26], the estimated disparities or depths in such passive ways are not so convincing for explicit pixel warping. As a result, disparity errors give rise to significant artifacts such as tearing and ghosting, especially in the occluded areas and along object edges.

To mitigate the dependency of explicit depth or disparity information for LFSR, we propose an Implicitly Multi-scale Fusion (IMsF) scheme to accumulate contextual information from multiple scales of the same image patch. In the absence of explicit depth information, we facilitate a set of sequent filters to convolve the same location to aggregate short- and long-range contextual information essential for Super-Resolution (SR) reconstruction. The feature maps obtained by each filter are rescaled before adding them up to obtain multi-scale encoding feature representation. With the help of IMsF layer, the network itself will pay attention to the most useful contextual information for SR reconstruction, without stiffly inferring the disparities for pixel warping.

Manuscript received July 18, 2017; revised December 5, 2017 and February 5, 2018; accepted April 26, 2018. Date of publication May 9, 2018; date of current version June 1, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61427811 and Grant 61573360, and in part by the National Key Research and Development Program of China under Grant 2016YFB1001000 and Grant 2017YFB0801900. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gustavo K. Rohde. *(Corresponding author: Zhenan Sun.)*

Y. Wang is with the University of Science and Technology of China, Hefei 230027, China, and also with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yunlong.wang@cripac.ia.ac.cn).

F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: fei.liu@nlpr.ia.ac.cn; kunbo.zhang@ia.ac.cn; gqhou@nlpr.ia.ac.cn; znsun@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2834819

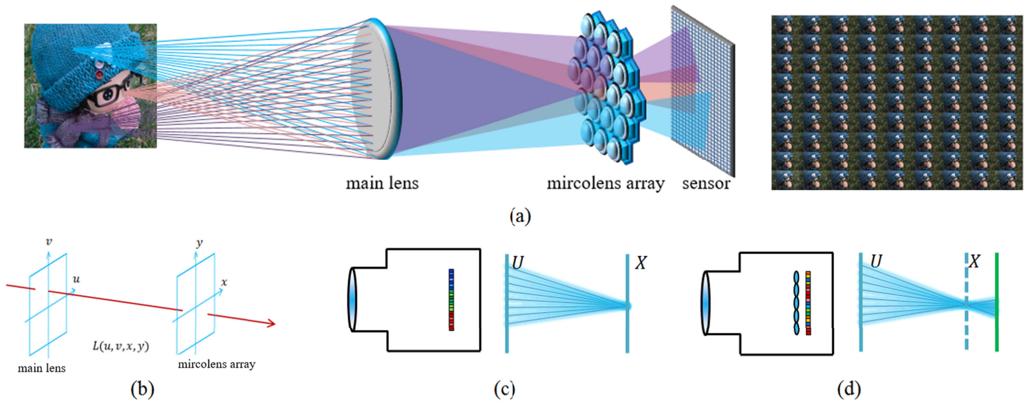


Fig. 1. LF Imaging. (a) Schematic structure of a microlens-based LF camera. A raw LF image can be decoded into a set of sub-aperture images on a 2D grid. (b) Two-plane parameterization model. (c) Basic structure of a conventional integral camera. (d) Basic structure of a LF camera inserted with microlens enabling to capture directions of light rays.

With the emergence of large LF datasets [27], data-driven learning methods based on deep neural network models have been successfully applied to LFSR. Yoon *et al.* [28], [29] are the first to apply the CNN framework to the research of LFSR. They propose a deep-learning structure called LFCNN composed of a spatial SR network and an angular SR network to jointly increase the spatial and angular resolution. However, in LFCNN, only pairs of sub-aperture images are fed to the convolutional network without exploiting the high correlation among adjacent views. Given the fact that Recurrent Neural Networks (RNN) [30] can well model long-term correspondences for temporal sequences, we incorporate IMsF layers into Bidirectional Recurrent Convolutional Neural Network (BRCNN) structure to iteratively model the spatial relations between two adjacent sub-aperture images forward and backward. Within the network, the common recurrent convolutions utilized by RNN [30] are modified to be more effective and flexible for LFSR. We also propose a Horizontal BRCNN (H-BRCNN) model and a Vertical BRCNN (V-BRCNN) model to respectively model the spatial correlations horizontally and vertically. Stacked generalization technique [31] is employed to ensemble these two models for final high-resolution predictions. All of these elements constitute the proposed framework LFNet (Fig.2).

The main contributions of this paper are as follows.

- Bidirectional Recurrent Convolutional Neural Network embedded with Implicitly Multi-scale Fusion layers is proposed to iteratively model the spatial correlations between two adjacent sub-aperture images of LF data. Within the structure, the recurrent convolutions are modified to be more effective and flexible for LFSR.
- Two sub-networks of the same structure are built to respectively model spatial correlations between neighbouring views horizontally and vertically, then stacked generalization is employed to ensemble these two models for final outputs.

The remainder of this paper is organised as follows. In Section 2, related work is introduced. Details of the method are studied in Section 3. Experimental results are presented in Section 4. Finally, we conclude our paper in Section 5.

II. Related Work

A. LF Imaging

Photographs taken by cameras are projections of the high-dimensional light signal onto the sensor plane. The microlens-based LF camera is able to record 4D light fields [4] via inserting a microlens array in front of its image sensor (Fig.1(a)). The prototype of the plenoptic camera was firstly introduced by Adelson and Wang [2] to infer depth from single lens in 1992. Ng *et al.* [5] design the hand-held LF camera (*aka.* Plenoptic 1.0) along with the imaging analysis and the digital refocusing method in 2005. On account of the low resolution limit, Lumsdaine and Georgiev [6] propose the focused plenoptic camera (*aka.* Plenoptic 2.0) by setting the microlens array focused at the focal plane of the main lens but not the main lens. In this paper, we concern about images captured by LF cameras with Plenoptic 1.0 optical structure.

Compared with conventional cameras (Fig.1(c,d)), the raw LF image comprises a large number of sub-aperture images formed by gathering the pixels of the same position in the coordinates covered by each microlens. The set of sub-aperture images are equivalently captured by a pinhole camera array settled at the aperture plane with slightly shifted views. Two-plane parameterization model [5] (Fig.1(b)) is usually applied to represent 4D light fields. Each light ray

$$L_F(x, y, u, v)$$

is illustrated by the interactions with two parallel planes, travelling from the coordinate $\mathbf{u} = (u, v)'$ (apostrophe denotes transpose) on the lens plane to the coordinate $\mathbf{x} = (x, y)'$ on the microlens array plane as Fig.1(b). By fixing one angular dimension (for instance $i.e.v = v^*$), 4D LF data $L_F(x, y, u, v)$ is dimensionally reduced to 3D LF data $L_F(x, y, u)$ as

$$L_F(x, y, u) = L_F(x, y, u, v^*)$$

Bolles *et al.* [32] analyze from the perspective of motion that a group of Epipolar-Plane Images (EPI) can be reorganized as a clip of frame sequence. Moreover, EPI are captured under the condition of nearly static scene, unchanged field of view and constant perspective transformation. By analogy, we argue that sub-aperture images of 3D light fields also have such simple structure by regular spacing of shifted views. In contrast with common videos, the spatial continuity and dependency between adjacent sub-aperture images of 3D LF data are much more regular. We exploit this unique clue of LF data by modeling the spatial correspondences between adjacent sub-aperture images for LFSR.

B. Light-Field Image Super-Resolution

To handle the resolution trade-off of LF cameras, most of the preceding methods can be divided into two main categories, *i.e.* reconstruction-based frameworks and learning-based algorithms. Reconstruction-based methods require both accurate geometric information of the scene as priors and fine parameter tuning for optimization, which are usually computational expensive. Bishop *et al.* [15] design a Bayesian framework to recover more information from geometric structure of the scene and super-resolve LF image. Lim *et al.* [17] analyze that 2D angular resolution contains spatially subpixel shifted information, which provides the redundant data used by SR algorithms. Georgiev and Lumsdaine [16] also establish subpixel correspondences with registration provided by the geometry of the microlens array. After disparity estimation based on EPI, Wanner and Goldluecke [19] optimize a variational framework to generate super-resolved novel views of the scene.

Other reconstruction-based works for LFSR basically focus on projection and resample of LF data [27], [33]–[38]. Liang and Ramamoorthi [20] demonstrate that typical LF cameras preserve frequency components above the spatial Nyquist rate and achieve spatial resolution above the microlens resolution with depth information as guidance to project the LF samples. Wang *et al.* [39] redefine one mapping function between the disparity of certain pixel and its shearing shift which relieves the dependency of camera parameters and depth information in the projection-based methods. Cho *et al.* [40] describe the procedures to calibrate a raw Lytro image and propose a dictionary-learning based interpolation method for LFSR. Marwah *et al.* [41] introduce light field atoms and utilize overcomplete dictionaries as a sparse representation of natural light fields for compressive LF photography. Farrugia *et al.* [42] propose a dictionary-learning based method for LFSR which learns the mapping between low-resolution (LR) and high-resolution (LR) patch volumes.

Projection-based SR methods for LF image are closely related to focal stack rendering. Using a plenoptic camera, conventional photographs focused on certain planes can be obtained through specific projections of the 4D light fields onto two spatial dimensions [5], [33]. Light fields and focal stacks are composed of multiple images that are either seen through different portions of the aperture, or focused at varying depths [38], [43]. The focal stack transform introduced by Nava *et al.* [34],and Jacobs *et al.* [35] could estimate the all-in-focus image of a scene at high resolution. Pérez *et al.* [36], [37] propose the fourier slice super-resolution to get the super-resolved discrete focal stack transform. Lee and Tai [44] state that the correlation among differently focused narrow depth-of-field images in a focal stack can be used to infer HR details for SR. It should be noted that these works that deal with focal stack transformation are quite different from the problem discussed in this paper, which jointly super-resolves sub-aperture images of LF data.

Limited by the small size of former LF datasets several years ago [45], learning-based methods for LFSR are rarely applicable. Yoon *et al.* [28] are the first to apply CNN framework to the research of LFSR. They propose a new deep learning structure called LFCNN composed of a spatial SR network and an angular SR network to jointly increase the spatial and angular resolution. In subsequent work [29], they refine the network architecture with a single spatial SR network and share some portions of the convolutions in angular SR network, which reduces the number of parameters by half. However, in LFCNN [28], [29], pairs of sub-aperture images are directly fed to the convolutional network without modeling the spatial correlations between them.

C. Deep-Learning Frameworks for Image Super-Resolution

Deep-learning methods [46]–[50] have been proven to make remarkable progress in modern vision tasks such as classification, detection, recognition, *etc*. Among the deep-learning methods for Single Image Super-Resolution (SISR), SRCNN is a representative state-of-the-art framework proposed by Dong *et al.* [51], which learns the mapping from LR to HR image in an end-to-end manner. SRCNN only consists of three layers, *i.e.* patch representation, non-linear mapping and reconstruction. Based on this work, Dong *et al.* [52] re-design the SRCNN structure to achieve a speed up of more than 40 times with even superior restoration quality named FSRCNN. Compared with SRCNN's shallow models, Kim *et al.* [53] propose VDSR, which is a very deep network with 20 weight layers for SISR. They use residual-learning and high learning rates to optimize the network at a fast speed of convergence. Also, Kim *et al.* [54] propose DRCN structure with a very deep recursive layer, which can improve SR results without introducing new parameters. Tai *et al.* [55] combine residual learning and recursive learning to build a deeper CNN model called DRRN. More recently, Lai *et al.* [56] propose LapSRN which adopts the Laplacian pyramid to progressively reconstruct the sub-band residuals of high-resolution images.

D. Multi-Frame Super-Resoltuion

Multi-frame SR techniques generally exploit the subpixel shift between input images to achieve spatial resolution enhancements. A large portion of the literature concern about VideoSR, which mainly model and exploit temporal correspondences among video frames. Baker and Kanade [57] extract optical flow to model the temporal correspondences in video sequences for video SR. Then, various improvements [58], [59] are explored to better handle visual motions. However, these methods suffer from the high computational cost and low accuracy due to the motion estimation. Moreover, motion estimation and global warping models of video SR frameworks are not suitable for direct application of spatial SR

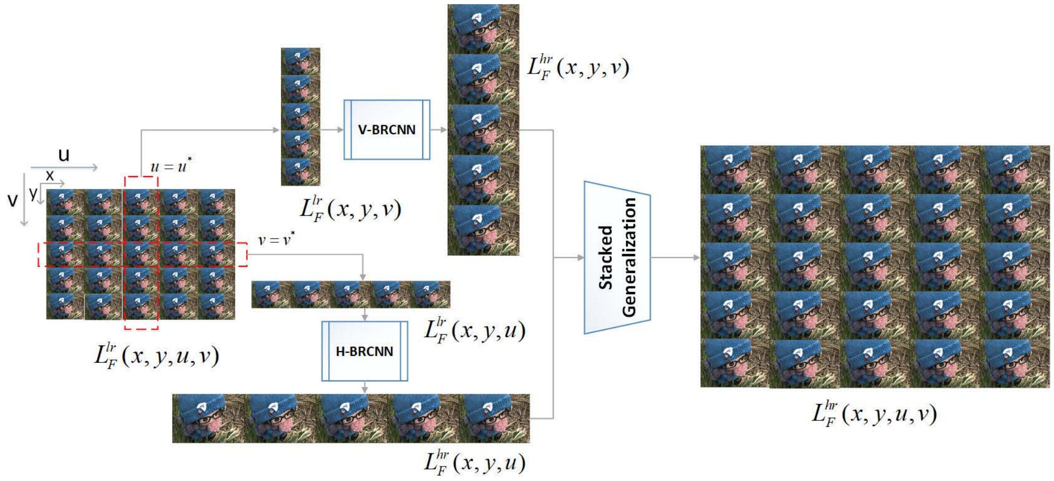


Fig. 2. Overview of the proposed framework *LFNet*. The degraded low-resolution 4D LF data $L_F^{lr}(x, y, u, v)$ as input is transformed to horizontal $3D$ LF data $L_F^{lr}(x, y, u)$ by setting $v = v^*$ and vertical $3D$ LF data $L_F^{lr}(x, y, v)$ by setting $u = u^*$. The spatial resolution of $L_F^{lr}(x, y, u)$ is enhanced by H-BRCNN and $L_F^{lr}(x, y, v)$ is enhanced by V-BRCNN. Once predictions of $L_F^{hr}(x, y, u, v^*)$ and $L_F^{hr}(x, y, u^*, v)$ are acquired, stacked generalization technique is employed to ensemble them for the final output $L_F^{hr}(x, y, u, v)$.

in LF data. Huang *et al.* [30] propose a Bidirectional Recurrent Convolutional Network (BRCN) as an end-to-end framework to efficiently learn the temporal correspondences for multi-frame SR, which achieves better performance and faster speed. Recently, Caballero *et al.* [60] exploit temporal correlations and improve reconstruction accuracy while maintaining real-time speed by introducing spatio-temporal subpixel convolution networks. The category of video SR approaches appears to be similar to LFSR, but not specially tailored for LF structure.

III. Methodology

A. Overview

Given degraded low-resolution $4D$ LF data $L_F^{lr}(x, y, u, v)$ at the resolution of (H, W, P, P), the goal of LFSR is to restore its high-resolution counterpart $L_F^{hr}(x, y, u, v)$ at the resolution of (sH, sW, tP, tP) with recovered high-frequency details,

$$L_F^{lr}(x, y, u, v) \xrightarrow{LFSR} L_F^{hr}(x, y, u, v)$$

where s is the upsampling factor for spatial dimensions while t for angular dimensions. In this paper, we focus on upsampling the spatial resolution (x, y) while not the angular resolution (u, v) of $4D$ LF data, *i.e.* $s > 1, t = 1$. One sub-aperture image from the view (i, j) of $L_F^{lr}(x, y, u, v)$ can be represented as

$$I_{i,j}^{lr} = L_F^{lr}(x, y, i, j)$$

$L_F^{lr}(x, y, u, v)$ is transformed to horizontal $3D$ LF data

$$L_F^{lr}(x, y, u) = L_F^{lr}(x, y, u, v^*)$$

by setting $v = v^*$ and vertical $3D$ LF data by setting $u = u^*$.

$$L_F^{lr}(x, y, v) = L_F^{lr}(x, y, u^*, v)$$

If the angular resolution of $L_F^{lr}(x, y, u, v)$ is $P \times P$, $3D$ LF data consist of P sub-aperture images

$$I_p^u = L_F^{lr}(x, y, u_p)$$

or

$$I_p^v = L_F^{lr}(x, y, v_p)$$

where I_p^u and I_{p+1}^u (or I_p^v and I_{p+1}^v, $p \in \{1, 2, \ldots, P-1\}$) are sub-aperture images from neighbouring views, which have nearly constant perspective transformation. Hence, we treat $L_F^{lr}(x, y, u)$ and $L_F^{lr}(x, y, v)$ as a sequence of N images to respectively enhance the spatial resolution of $3D$ LF data horizontally and vertically. Once predictions of $L_F^{hr}(x, y, u, v^*)$ and $L_F^{hr}(x, y, u^*, v)$ are acquired, stacked generalization technique [31] is employed to ensemble them for the final output $L_F^{hr}(x, y, u, v)$ as shown in Fig.2.

B. Implicitly Multi-Scale Fusion Layer

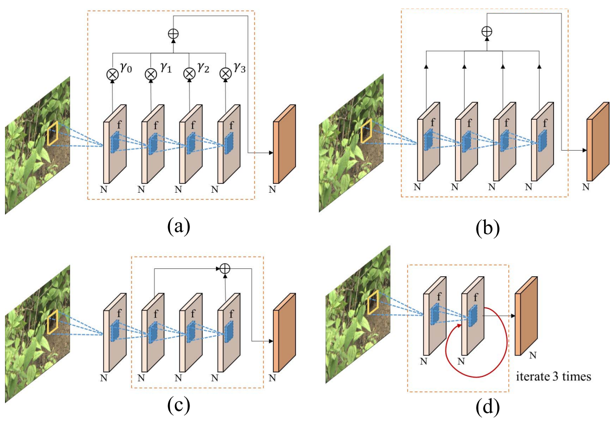
Multi-scale inference has been utilized in GoogLeNet's *Inception* structure [48] to effectively aggregate local information, allowing more robust and accurate predictions. More recently, Lai *et al.* [56] adopts the idea of Laplacian pyramid in SISR which facilitates multi-scale information. Considering that the disparities between two adjacent sub-aperture images are generally around a few pixels according to the scene, small areas centered at the same location affect pixel warping when explicitly registered. Moreover, contextual information from multiple scales of the image patch contains short- and long-range essential elements for SR reconstruction.

As shown in Fig.3, we propose IMsF layer by facilitating four sequent convolutional layers with the same kernel size f and channel N to convolve the same image region as

$$h_1 = \sigma(W_1 * I_{in} + b_1) \quad (1)$$

$$h_n = \sigma(W_n * h_{n-1} + b_n), n = 2, 3, 4 \quad (2)$$

where I_{in} denotes an input view in $3D$ LF data, W_n and b_n are the convolution kernel and bias at layer n, operator $*$ represents convolutional operation, $\sigma(\cdot)$ is the activation function. The size of W_1 is $N \times c \times f \times f$ and $W_k (k \in \{2, 3, 4\})$



(a) (b) (c) (d)

Fig. 3. Comparisons of an IMsF layer to other structures. (a) Proposed IMsF layer. Four sequent convolutional layers of the same kernel size f and channel number N are facilitated to convolve the same region from the reference view. The activations of each layer are rescaled by learned weights, then added together to obtain multi-scale encoding feature representation. (b) IMsF without rescaling. The activations of each layer are directly added up. (c) Resnet Blocks [50]. (d) DRCN block [54].

are $N \times N \times f \times f$, where N is the number of filters, c is the number of input channels, f is the filter size. A straightforward way to fuse h_n is to directly add them up. However, we experimentally find that adding them up directly would greatly weaken the representative ability of the model (Section IV.B). Rather, we introduce additional parameters γ_n to rescale h_n as Eq.3, where operator $\otimes$ denotes the channel-wise multiplication. These parameters are learned along with the other model parameters, and effectively strengthen the representation power of the network.

$$\widehat{h_n} = \gamma_n \otimes h_n \tag{3}$$

The rescaled representation $\widehat{h_n}$ are then added together as Eq.4. H_{ms} denotes the multi-scale encoded representation and b_{ms} is the bias.

$$H_{ms} = \sum_{n \in \{1,2,3,4\}} \widehat{h_n} + b_{ms} \tag{4}$$

With the multi-context feature representation encoded by IMsF layer, the network itself will adaptively focus on the contextual information that is the most useful for accurately upsampling. Through IMsF, we pay more attention on contextual information fusion from multiple scales rather than explicit pixel warping or image registration.

C. Bidirectional Recurrent Convolutional Neural Network

A group of sub-aperture images in $3D$ LF data can be reorganized as a clip of frame sequences with static scene, unchanged field of view and constant perspective transformation. In the absence of explicit depth information as priors, we incorporate IMsF layers into BRCNN structure to iteratively model the spatial relations between neighbouring views. The proposed IMsF-BRCNN is composed of a forward sub-network and a backward sub-network as shown in Fig. 4. In order to enhance the spatial resolution of $3D$ LF data, we propose H-BRCNN and V-BRCNN to respectively model the horizontal and vertical spatial relations, which are of the same network structure. To alleviate the border effect of BRCNN, 3D LF data are padded with I_0 the same as I_1 in the forward sub-network. In the backward network, 3D LF data are padded with I_{N+1} the same as I_N as Fig.4.

In the first step, each pair of sub-aperture images denoted by I_o (one sub-aperture image) and I_{adj} (adjacent to I_o) in $3D$ LF data[1] is fed to IMsF layer to get the multi-scale encoded feature representation, H^o_{ms} for I_o and H^{adj}_{ms} for I_{adj} respectively. Note that all the IMsF layers in BRCNN share the parameters.

To model the spatial correlations between neighbouring views, two hidden layers for reconstruction with recurrent convolutions are introduced. In the first hidden layer, activations for adjacent sub-aperture image of this layer H^{adj}_1 and H^o_{ms} are considered as inputs to get H^o_1 in Eq.5.

$$H^o_1 = \sigma(W_{r1} * cat(W_{c1} * H^o_{ms}, H^{adj}_1) + b_1) \tag{5}$$

where W_{r1} denotes the kernel of recurrent convolution, W_{c1} is the kernel of the feedforward convolution for the first layer, b_1 is the bias, $cat(\cdot,\cdot)$ represents concatenating operation along the channel dimension as shown in Fig.5(b). The size of W_{r1} is $N_1 \times 2N_1 \times 1 \times 1$ and that of W_{c1} is $N_1 \times N \times f_1 \times f_1$, where N is channel number of H^o_{ms}, N_1 and f_1 are the filter number and size of W_{c1}. It can be deduced that recurrent convolutions utilized in BRCN by Huang *et al.* [30] as Eq.6 can be regarded as a special case of our structure,

$$H^o_1 = \sigma(W'_{c1} * H^o_{ms} + W'_{r1} * H^{adj}_1 + b_1) \tag{6}$$

in case that the first half parameters $N_1 \times N_1 \times 1 \times 1$ of W_{r1} for weighting $W_{c1} * H^o_{ms}$ would be fixed to 1 in entries of $(n_1, n_1, 1, 1)$, where $n_1 \in \{0, 1, \ldots, N_1\}$ and others would be set to 0. Meanwhile, those of the second half $N_1 \times N_1 \times 1 \times 1$ for weighting H^{adj}_1 are the same as BRCN [30]. In our modified design of recurrent convolutions, parameters for weighting $W_{c1} * H^o_{ms}$ are also learnable rather than fixed to 1, which is more effective and flexible in modeling the spatial correlations between neighbouring views. Within the forward sub-network,

$$H^{n+1}_{f1} = \sigma(W^f_{r1} * cat(W^f_{c1} * H^{n+1}_{ms}, H^n_{f1}) + b^f_1) \tag{7}$$

where $n \in \{0, 1, \ldots, N-1\}$ and f denotes $forward$. Within the backward sub-network,

$$H^n_{b1} = \sigma(W^b_{r1} * cat(W^b_{c1} * H^n_{ms}, H^{n+1}_{b1}) + b^b_1) \tag{8}$$

where $n \in \{1, 2, \ldots, N\}$ and b denotes $backward$.

In the second hidden layer, the inputs are H^o_1 and H^{adj}_2 to produce H^o_2,

$$H^o_2 = \sigma(W_{r2} * cat(W_{c2} * H^o_1, H^{adj}_2) + b_2) \tag{9}$$

where W_{r2} and W_{c2} denote convolutional kernels of recurrent and feedforward convolutions in the second layer, b_2 is the bias. The size of W_{r2} is $N_2 \times 2N_2 \times 1 \times 1$ and that of W_{c2} is $N_2 \times N_1 \times f_2 \times f_2$, where N_1 is channel number of H^o_1, N_2

[1] For clarity, $horizontal$ and $vertical$ are not differentiated here.

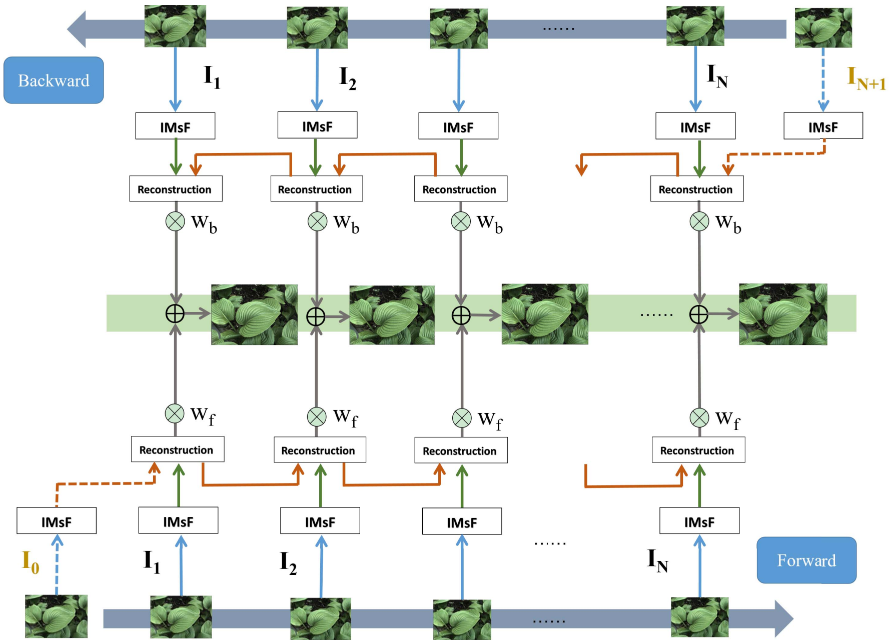


Fig. 4. Structure of the proposed IMsF-BRCNN. 3*D* LF data containing a sequence of *N* views is fed into the network. The network structure consists of a *forward* sub-network and a *backward* sub-network to model the spatial correlations between two neighbouring views from different directions. The two sub-networks share the parameters of all IMsF layers while other layers don't. w^f and w^b are learnable parameters to get weighted average of both predictions.

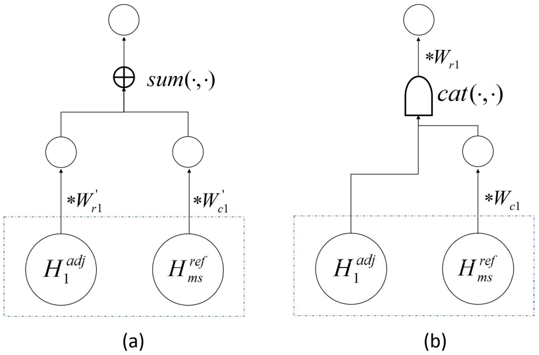


Fig. 5. Comparisons of recurrent convolutions. (a) The structure utilized by Huang *et al.* [30] (b) The proposed recurrent convolutions tailored for LF structure. (a) can be regarded as a special case of (b).

and f_2 are filter number and kernel size of W_{c2}. Within the forward sub-network,

$$H_{f2}^{n+1} = \sigma(W_{r2}^f * cat(W_{c2}^f * H_{f1}^{n+1}, H_{f2}^n) + b_2^f) \qquad (10)$$

within the backward sub-network,

$$H_{b2}^n = \sigma(W_{r2}^b * cat(W_{c2}^b * H_{b1}^n, H_{b2}^{n+1}) + b_2^b) \qquad (11)$$

In the output layer, we just carry out feedforward convolutions to yield the estimated HR predictions I_{est}^o without the activation function,

$$I_{est}^o = W_{out} * H_2^o + b_{out} \qquad (12)$$

where W_{out} denotes the kernel of feedforward convolution in the output layer and b_{out} the bias. The size of W_{out} is $c \times N_2 \times f_{out} \times f_{out}$, where N_2 is channel number of H_2^o, c and f_{out} are output channel and filter size of W_{out}. Within the forward sub-network,

$$I_n^f = W_{out}^f * H_{f2}^n + b_{out}^f \qquad (13)$$

within the backward sub-network,

$$I_n^b = W_{out}^b * H_{b2}^n + b_{out}^b \qquad (14)$$

Further, we introduce learnable parameters w^f and w^b to get weighted average of these two predictions,

$$I_{est}^n = w^f I_n^f + w^b I_n^b \qquad (15)$$

D. Loss Function and Stacked Generalization

Most of the preceding deep-learning-based methods for SR perform network learning and parameters updating by minimizing the MSE loss function between the predicted HR outputs I_{pred} and the ground truth I_{gt} via Stochastic Gradient Descent (SGD).

$$L_{MSE} = \frac{\left\| I_{gt} - I_{pred} \right\|_2^2}{numel(I_{gt})} \qquad (16)$$

where $\|\cdot\|_2$ denotes L_2 Norm, $numel(\cdot)$ denotes number of elements. The weight decay parameter is also popular in governing the regularization term of the neural network to avoid overfitting. Hence, the unified loss function of our network to minimize during training combines MSE loss function L_{MSE} and L_1 norm of the net parameter Θ as Eq.17, where λ is the balancing parameter of weight decay.

$$L = L_{MSE} + \lambda \|\Theta\|_1 \qquad (17)$$

By using pre-trained network model of H-BRCNN for inference, it yields HR predictions I_{est}^{H} in horizonal $3D$ LF data $L_{F}^{lr}(x, y, u)$,

$$I_{est}^{H} = \Psi(L_{F}^{lr}(x, y, u); \theta_H) \tag{18}$$

where θ_H denotes the network parameters of H-BRCNN. By analogy, V-BRCNN produces HR predictions I_{est}^{V} in vertical $3D$ LF data $L_{F}^{lf}(x, y, v)$,

$$I_{est}^{V} = \Psi(L_{F}^{lr}(x, y, v); \theta_V) \tag{19}$$

Stacked generalization proposed by Wolpert [31] is a way of combining multiple models. In this paper, we employ simple stacked generalization to ensemble the set of predictions $\{I_{est}^{H}, I_{est}^{V}\}$ through a linear combination as Eq.20.

$$I_{est} = w_H I_{est}^{H} + w_V I_{est}^{V} \tag{20}$$

w_H and w_V are optimized through gradient descent by minimizing the L_2 norm loss between ground truth I_{gt} and I_{est} as Eq.21.

$$\begin{aligned} &\underset{w_H, w_V}{\arg\min} \ \left\| I_{gt} - I_{est} \right\|_2 \\ &s.t. \ w_H + w_V = 1 \end{aligned} \tag{21}$$

E. Implementation Details

Within IMsF layers, we experimentally set $N = 64$, $f = 3$ and use zero padding to avoid border effects. The activation function $\sigma(\cdot)$ is rectified linear unit ($ReLu$), *i.e.* $\sigma(x) = \max(0, x)$. The filters of W_1 are randomly initialized from a zero-mean Gaussian distribution with standard deviation 0.01 and W_k ($k \in \{2, 3, 4\}$) are initialized in the same manner but with standard deviation 0.001, all the bias b_k ($k \in \{1, 2, 3, 4\}$) are initialized to zero. γ_k ($k \in \{2, 3, 4\}$) is initialized to 1. We only deal with luminance channel in the YCrCb color space so that the input channel c is 1.

In the proposed IMsF-BRCNN, we experimentally set $N_1 = 32$, $f_1 = 5$, $N_2 = 16$, $f_2 = 1$, $f_{out} = 9$. All the filters of convolutional layers are initialized from a zero-mean Gaussian distribution with standard deviation 0.01 except that the first half parameters of W_{r1} and W_{r2} are set to 1 in some certain entries. All the bias are initialized to 0.1. w^f and w^b are initialized to 0.5. Zero padding is also applied to avoid border effects. The forward and backward sub-network of the proposed IMsF-BRCNN don't share the parameters except the IMsF layer. There are two IMsF-BRCNNs in the proposed framework. Each has two sets of parameter, where θ_H^f and θ_H^b are updated for H-BRCNN, along with θ_V^f and θ_V^b updated for V-BRCNN.

The optimization is conducted by the mini-batch momentum SGD method with a batch size of 64, momentum of 0.9 and weight decay of $\lambda = 0.005$. The learning rate is initially set to $1e-3$ for the weights in the output layer while $1e-2$ for other layers. We then decrease the learning rate by a factor of 0.1 every 10 epochs until the validation loss converges.

The proposed model is implemented using the Theano package [61] and proceeded on a workstation with an Intel 3.6 GHz CPU and a TiTan X GPU. A LF image at the resolution of $625 \times 434 \times 7 \times 7$ can be $4\times$ spatially super-resolved within 4 seconds, roughly 0.08s per sub-aperture image. The source code and real-world datasets to reproduce the experimental results will be released upon the acceptance of submission.

IV. Experiments

A. Setup

To validate the effectiveness of the proposed framework, we conduct experiments on both public and self-captured LF datasets of synthetic and real-world scenes.

1) Synthetic Dataset: For fair comparison, we use synthetic LF images from the public HCI database [45] and follow the same protocol in LFCNN [28] to split the training and test datasets. $Buddha$ and $Mona$ are selected as test samples and the rest 10 LF samples are used to generate patches for training. Both sub-aperture images and ground-truth depths for all views of HCI synthetic datasets are provided at the spatial resolution of 768×768, angular resolution of 9×9. 3D LF patches of size $48 \times 48 \times 5$ are randomly cropped from the same region of 5 adjacent sub-aperture images horizontally or vertically in the training datasets as the ground truth HR samples. They are spatially $\times 2$ downsampled to $24 \times 24 \times 5$ and upsampled again using the bicubic interpolation to generate the corresponding LR training samples. We double the training datasets by adding a copy of each LF sample with permuted spatial dimensions, *i.e.* (x, y, u) to (y, x, u) and (x, y, v) to (y, x, v). Two training datasets are constructed in this manner for training H-BRCNN and V-BRCNN respectively, either of which has nearly 100000 pairs of LR and HR samples. The test data are also spatially $\times 2$ downsampled and re-upsampled using bicubic interpolation.

2) Real-World Dataset: To train the proposed network for real-world scenes, we take more than 200 LF images with a Lytro Illum camera, which include various light conditions, textures and depths under indoor and outdoor environment. We use Light Field Toolbox v0.4 [63] to decode the raw LF images and extract $4D$ LF data at the spatial resolution of 625×434 and the angular resolution of 9×9. In order to increase the difficulty of the task for real-world scenes, we follow the same protocol in BRCN [30] to generate the training and test datasets, *i.e.* 1) using the Gaussian filter with standard deviation 2 to smooth each sub-aperture image, 2) downsampling sub-aperture images by a factor of 4 which is usually considered as the most difficult case in SR. Similarly, we extract the training datasets for H-BRCNN and V-BRCNN in the same manner as stated in the synthetic experiments.

B. Evaluation of IMsF layer

To evaluate the effectiveness of proposed IMsF layer, we substitute this layer in LFNet with other structures, including IMsF without rescaling, Resnet block [50], DRCN block [54] as Fig.3. LFNet without IMsF layer serves as a baseline. Training on the same real-world datasets split with batch size 64, we record average PSNR on validation set every 200 iterations within 2×10^4 iterations. The results are depicted in Fig.6.

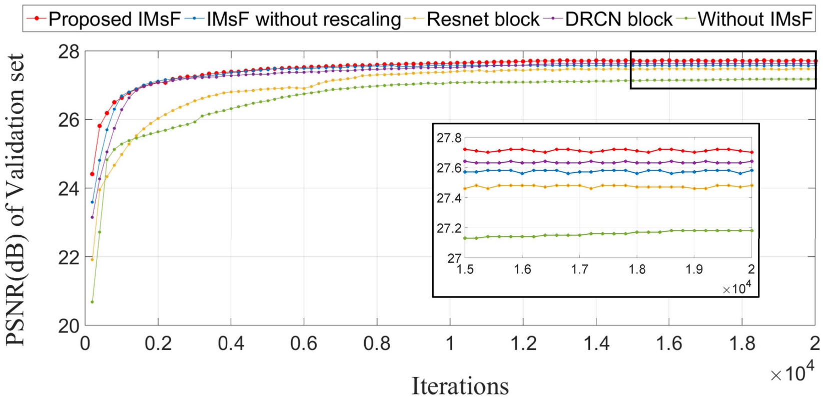


Fig. 6. Evaluation of IMsF layer. The proposed IMsF layer in LFNet is substituted with other structures, including IMsF without rescaling, Resnet block [50], DRCN block [54]. The PSNR of proposed IMsF layer after convergence is nearly 0.2dB higher than that of IMsF without rescaling and 0.5dB higher than the baseline, which can demonstrate that the proposed IMsF layer is effective in improving the performance of LFSR. Compared with other structures, the proposed IMsF layer achieves superior results with faster convergence speed.

TABLE I
QUANTITATIVE EVALUATIONS OF HCI SYNTHETIC LF DATASETS

Methods	PSNR(dB)						SSIM					
	Buddha			Mona			Buddha			Mona		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Bicubic	34.12	34.52	34.98	34.21	34.48	34.67	0.9238	0.9325	0.9458	0.9241	0.9321	0.9451
Mitra [62]	28.83	29.91	31.17	28.54	29.28	30.07	0.8665	0.8994	0.9343	0.8541	0.8911	0.9319
Wanner [19]	24.43	29.69	36.97	25.40	30.76	37.60	0.7662	0.8691	0.9470	0.8542	0.9324	0.9862
Wang [39]	36.21	36.21	36.21	36.73	36.73	36.73	0.9571	0.9571	0.9571	0.9605	0.9605	0.9605
Yoon [29]	36.25	36.95	37.35	37.03	37.99	38.53	0.9579	0.9623	0.9657	0.9833	0.9863	0.9878
H-BRCNN	38.08	38.37	38.71	**38.39**	38.72	38.81	0.9701	0.9722	0.9752	0.9881	0.9890	0.9892
V-BRCNN	**38.10**	38.39	38.76	38.37	**38.74**	**38.82**	0.9705	0.9726	0.9755	0.9882	0.9889	0.9893
LFNet	38.09	**38.42**	**38.77**	38.38	38.73	38.80	**0.9709**	**0.9731**	**0.9760**	**0.9884**	**0.9891**	**0.9895**

[1] Bold texts indicate the best results. Underlined texts indicate second best results.

Compared with other structures, the proposed IMsF layer achieves superior results with faster convergence speed. The DRCN block [54] obtains a little higher PSNR than IMsF layer without rescaling but converges a bit more slowly. The PSNR of proposed IMsF layer after convergence is nearly 0.2dB higher than that of IMsF without rescaling and 0.5dB higher than the baseline, which can demonstrate that the proposed IMsF layer is effective in improving the performance of LFSR.

C. Quantitative Evaluations

1) Results on Synthetic Dataset: In synthetic experiments, we compare LFNet with four LFSR methods [19], [29], [39], [62]. Table I shows the results of quantitative evaluations on HCI synthetic datasets. The methods in [19], [39], and [62] are all reconstruction-based frameworks that require disparity or depth information as priors to independently super-resolve each view of LF data. As HCI synthetic datasets provide ground-truth depths, we just downsample the required disparity maps to the resolution of LR sub-aperture images accordingly. LFCNN [29] is a deep-learning-based structure which contains a spatial SR network and an angular SR network. We only compare with its spatial SR network. The results of these comparative methods are obtained via the source codes provided by the authors and the parameters are carefully tuned to maximize performance.

The peak signal-to-noise ratio (PSNR) and the gray-scale structural similarity (SSIM) [64] are used as the performance indexes. We measure the PSNR and SSIM scores of all the estimated HR views and report the minimum (*Min*), average (*Avg*) and maximum (*Max*) values for each LF sample. As seen in Table I, our method outperforms other methods by a large margin in both PSNR and SSIM. The proposed approach can surpass state-of-the-art LFCNN [29] by 1.47 dB of *Buddha* and 0.74 dB of *Mona* on average. Due to degradation of sub-aperture images and corresponding disparities, the performance of [19] and [62] drops sharply, even worse than simple bicubic interpolation.

2) Results on Real-World Dataset: For real-world scenes, we compare LFNet with seven methods: three LFSR methods [29], [39], [62], one video SR method BRCN in [30] and three SISR method, including FSRCNN in [52], VDSR in [53], DRRN in [55]. Per-view disparity map needed by the method in [39] is estimated based on EPI from 4D LF data. PSNR and SSIM are also chosen as performance indexes. We select 6 LF samples captured by Lytro Illum Camera for comparisons including *Stone*, *Bush*, *Glass*, *Door*, *Pillar*, *Flower*. The minimum (*Min*), average (*Avg*), and maximum (*Max*) values for each LF sample are reported.

TABLE II

QUANTITATIVE EVALUATIONS OF REAL-WORLD SCENES CAPTURED BY LYTRO ILLUM CAMERA

Methods	PSNR(dB)																	
	Stone			Bush			Glass			Door			Pillar			Flower		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Bicubic	26.90	27.01	27.13	26.81	26.91	27.20	30.50	30.57	30.70	27.52	27.61	28.07	28.40	28.54	28.96	28.42	28.63	29.00
Mitra [62]	25.26	25.75	26.34	24.94	25.18	25.34	26.41	26.73	26.89	25.26	25.80	26.34	26.20	26.32	26.44	24.31	24.46	24.67
Wang [39]	25.16	25.16	25.16	24.98	24.98	24.98	26.89	26.89	26.89	26.04	26.04	26.04	27.25	27.25	27.25	25.07	25.07	25.07
Yoon [29]	27.90	28.03	27.18	28.25	28.37	28.66	31.74	31.86	32.01	28.29	28.39	28.89	29.60	29.79	30.17	30.31	30.58	30.99
Dong [52]	27.96	28.10	28.25	28.45	28.59	28.88	32.13	32.21	32.32	28.49	28.59	29.10	29.93	30.07	30.55	30.50	30.75	31.13
Huang [30]	28.00	28.13	28.29	28.28	28.42	28.67	31.91	32.02	32.11	28.52	28.64	29.13	29.41	29.61	30.03	30.24	30.46	30.88
Kim [53]	28.09	28.25	28.37	28.74	28.89	29.15	32.69	32.78	32.93	28.65	28.77	29.24	30.20	30.28	31.01	30.75	31.02	31.43
Tai [55]	28.19	28.40	28.49	28.86	29.01	29.25	32.75	32.98	33.12	28.80	28.91	29.53	**30.32**	30.42	**31.14**	30.86	31.18	31.61
H-BRCNN	28.37	28.51	28.66	**29.03**	29.19	29.50	33.04	33.17	33.26	28.89	**29.04**	29.53	30.29	30.46	31.03	31.06	31.33	31.76
V-BRCNN	28.35	28.52	28.67	28.90	29.09	29.38	33.02	33.13	33.28	28.90	29.01	29.53	30.27	30.43	30.98	31.09	31.38	31.83
LFNet	**28.38**	**28.54**	**28.68**	29.02	**29.20**	**29.52**	**33.05**	**33.22**	**33.29**	**28.91**	29.03	**29.54**	30.30	**30.48**	31.01	**31.10**	**31.39**	**31.87**

Methods	SSIM																	
	Stone			Bush			Glass			Door			Pillar			Flower		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Bicubic	0.6274	0.6403	0.6542	0.7217	0.7247	0.7303	0.7546	0.7570	0.7644	0.7408	0.7444	0.7567	0.7478	0.7536	0.7700	0.8035	0.8138	0.8262
Mitra [62]	0.6517	0.6640	0.6776	0.6832	0.6958	0.7035	0.8040	0.8150	0.8158	0.7145	0.7379	0.7587	0.7229	0.7304	0.7433	0.6905	0.6997	0.7124
Wang [39]	0.6122	0.6122	0.6122	0.6197	0.6197	0.6197	0.7359	0.7359	0.7359	0.6628	0.6628	0.6628	0.6881	0.6881	0.6881	0.6498	0.6498	0.6498
Yoon [29]	0.6735	0.6864	0.6997	0.7774	0.7803	0.7859	0.8882	0.8912	0.8933	0.7738	0.7774	0.7885	0.7818	0.7879	0.8024	0.8445	0.8540	0.8649
Dong [52]	0.6762	0.6893	0.7029	0.7820	0.7850	0.7910	0.8895	0.8921	0.8941	0.7789	0.7826	0.7938	0.7840	0.7900	0.8044	0.8467	0.8556	0.8660
Huang [30]	0.6969	0.7094	0.7227	0.7849	0.7877	0.7943	0.8870	0.8899	0.8918	0.7869	0.7911	0.8013	0.7842	0.7906	0.8048	0.8501	0.8589	0.8694
Kim [53]	0.6864	0.6995	0.7189	0.7851	0.7911	0.7955	0.8906	0.8934	0.8950	0.7805	0.7838	0.7957	0.7861	0.7922	0.8065	0.8497	0.8592	0.8702
Tai [55]	0.6895	0.7044	0.7201	0.7886	0.7956	0.7987	0.8911	0.8951	0.8962	0.7831	0.7874	0.7998	0.7902	0.7961	0.8107	0.8536	0.8634	0.8741
H-BRCNN	0.7061	0.7187	0.7309	0.7961	0.7997	0.8058	0.8949	0.8981	0.8997	0.7953	0.7984	0.8075	0.7934	0.7988	0.8141	0.8617	0.8705	0.8804
V-BRCNN	0.7058	0.7186	0.7327	0.7976	0.8006	0.8073	0.8953	0.8984	0.9005	0.7948	0.7980	0.8084	0.7930	0.7991	0.8136	0.8629	0.8713	0.8809
LFNet	**0.7065**	**0.7193**	**0.7331**	**0.7984**	**0.8014**	**0.8075**	**0.8962**	**0.8994**	**0.9011**	**0.7962**	**0.7988**	**0.8098**	**0.7941**	**0.7998**	**0.8150**	**0.8637**	**0.8720**	**0.8819**

[1] Bold texts indicate the best results. Underlined texts indicate second best results.

Table II shows the quantitative evaluations of these LF samples. From Table II, it can be seen that the proposed approach yields higher PSNR and SSIM scores on average than other state-of-the-art methods among all the LF samples. It is quite challenging to estimate precise disparity maps from LF images of real-world scenes. So registration-based methods [39] and [62] fail to reconstruct reliable HR results, causing worse performances even than simple bicubic interpolation. The state-of-the-art learning-based SR methods aiming to super-resolve single image FSRCNN [52], VDSR [53], DRRN [55]; LF image LFCNN [29]; video frames BRCN [30] give quite close results. DRRN [55] achieves second best scores in all the samples on PSNR while BRCN [30] achieves second best scores in 2 out of 6 samples on SSIM. Our method exceeds second best results among these methods by 0.06 $\sim$ 0.24dB in PSNR and 0.0037 $\sim$ 0.0099 in SSIM, 0.16dB and 0.0068 on average.

As shown in Table I and II, H-BRCNN and V-BRCNN obtain very close results in PSNR. After ensembling them through stacked generalization, final SR results only achieve tiny improvements. However, nearly all the LF samples gain considerable increase in SSIM. It demonstrates that combination of H-BRCNN and V-BRCNN through stacked generalization makes SR results of the proposed framework closer to real image in structural similarity and better for human visual systems.

D. Qualitative Comparisons

Due to space limitation, we only display estimated HR central view of [30], [53], and [55] along with ground truth, bicubic interpolation and our method for qualitative comparisons. **Readers are strongly recommended to enlarge and view these figures on screen for better comparisons. See more results in the supplement.**

As shown in Fig.7-10, the results of VDSR [53] and DRRN [55] are oversmoothed with blurry details, although they achieve high PSNR and SSIM scores. BRCN [30] recovers more details but suffers from severe ringing artifacts especially along boundaries and edges. For instance, in Fig.7(e), textures on surface of the stones can't be clearly resolved and the boundary of the white line is rather vague. The result of BRCN [30] in Fig.9(e) reveals more plausible details of the door but comes along with severe ringing artifacts which are not visual-pleasing. By contrast, our SR results show photo-realistic details and much fewer ringing artifacts where boundaries of local structure are well preserved. The proposed framework LFNet obtains superior SR results than other state-of-the-art methods in visual effects.

E. Applications

As to the super-resolved LF image, adequate high-frequency details are restored after applying the proposed algorithm,

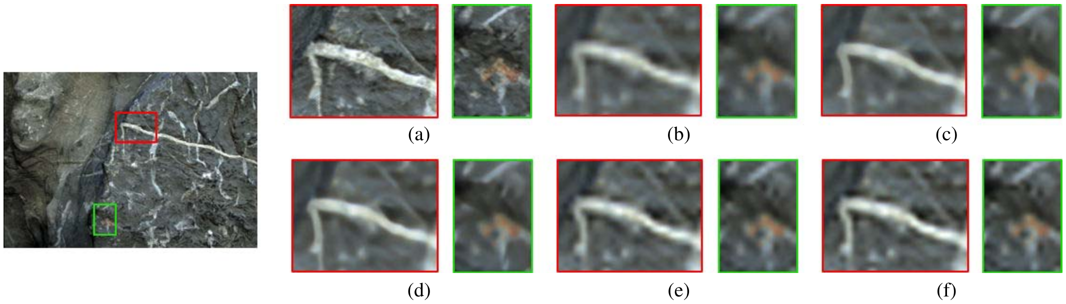


Fig. 7. Qualitative Comparisons of *Stone*. Only central view of 4D LF data is displayed. (a) Ground Truth (b) Bicubic Interpolation. (c) VDSR [53] (d) DRRN [55] (e) BRCN [30] (f) LFNet(ours).

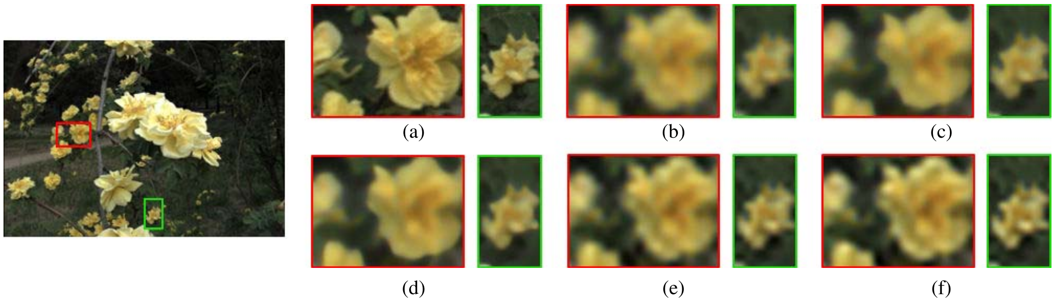


Fig. 8. Qualitative Comparisons of *Flower*. Only central view of 4D LF data is displayed. (a) Ground Truth. (b) Bicubic Interpolation. (c) VDSR [53] (d) DRRN [55] (e) BRCN [30] (f) LFNet(ours).

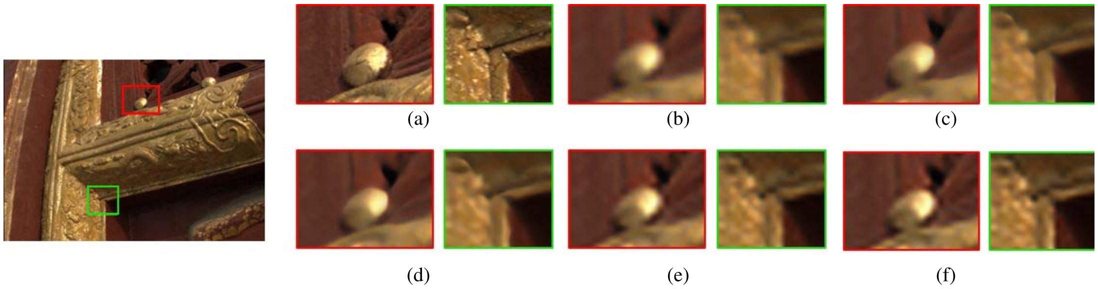


Fig. 9. Qualitative Comparisons of *Door*. Only central view of 4D LF data is displayed. (a) Ground Truth. (b) Bicubic Interpolation. (c) VDSR [53] (d) DRRN [55] (e) BRCN [30] (f) LFNet(ours).

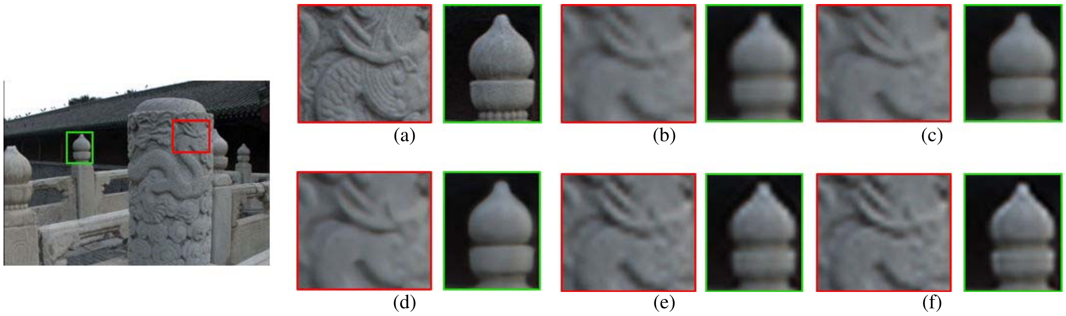


Fig. 10. Qualitative Comparisons of *Pillar*. Only central view of 4D LF data is displayed. (a) Ground Truth. (b) Bicubic Interpolation. (c) VDSR [53] (d) DRRN [55] (e) BRCN [30] (f) LFNet(ours).

which is favorably beneficial for real-world LF applications such as depth estimation. Table III shows quantitative comparisons on scenes *antinous* and *kitchen* from the 4D light field benchmark [65] by applying the depth estimation algorithm in [21] on SR results of bicubic interpolation, BRCN [30], VDSR [53], DRRN [55] and LFNet, along with ground truth. The selected samples are challenging with transparent reflections, occlusions and complex geometries. It can be seen that the proposed framework LFNet achieves better results than other methods.

To test the robustness of the proposed algorithm, we perform qualitative comparisons on real-world scenes. Taking LF sample *Bush* for instance, we apply the state-of-the-art depth estimation method [24] on ground-truth LF image,

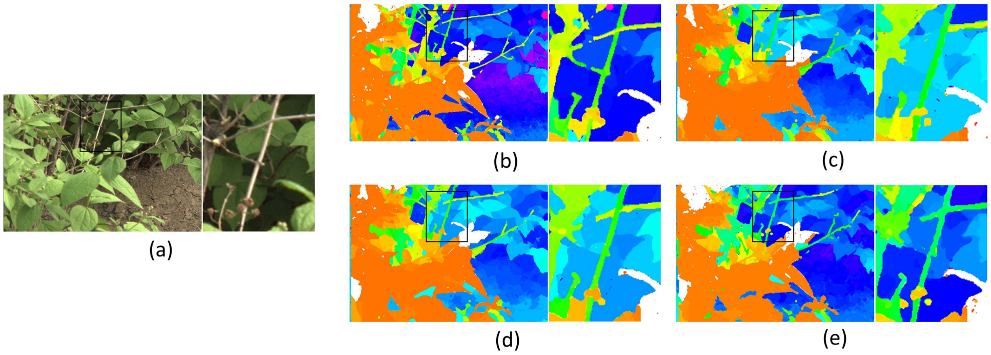


Fig. 11. The proposed method can enhance the performance of depth estimation. Warm color indicates nearby objects and cool color indicates distant objects. (a) Central View of LF sample *Bush*. (b) Estimated depth map from ground-truth LF image. (c) Estimated depth map from SR results of bicubic interpolation. (d) Estimated depth map from SR results of LFCNN [29]. (e) Estimated depth map from SR results of the proposed method.

TABLE III
RMSE (Root Mean Squared Errors) Statistics of Depth Estimation Application on HCI New Dataset [65]

Scenes	antinous	kitchen
Bicubic	0.1907	0.0822
BRCN [30]	0.1904	0.0813
VDSR [53]	0.1885	0.0818
DRRN [55]	0.1880	0.0811
GT	**0.1770**	**0.0721**
LFNet(ours)	0.1794	0.0810

SR results of bicubic interpolation, LFCNN [29] and the proposed approach. The estimated depth maps are shown in Fig.11. The depth map inferred from the upsampled LF image via our algorithm is closer to the result of ground-truth LF image with more accurate depth layers and structural boundaries, which demonstrates that the proposed method can enhance the performance of depth estimation.

V. Conclusion

In this paper, IMsF scheme has been proposed to accumulate contextual information from multiple scales for SR reconstruction. IMsF layers are then incorporated into BRCNN structure to iteratively model spatial correlations between two adjacent sub-aperture images of LF data. Within IMsF-BRCNN, the recurrent convolutions are specially tailored for LFSR. A horizontal and a vertical IMsF-BRCNN have been built to respectively super-resolve 3D LF data and ensembled through stacked generalization. We have validated the effectiveness of IMsF layer and the proposed method on both synthetic datasets and real-world scenes captured by a LF camera. Compared with state-of-the-art SISR, Video SR and LFSR methods, LFNet has achieved much better results not only in terms of PSNR and SSIM indexes but also of superior visual quality. Furthermore, we demonstrate that the proposed framework can enhance the performance of LF applications such as depth estimation. In the future, we will apply LFNet to more image restoration problems in LF imaging.

Acknowledgment

The authors would like to thank the associate editor and the reviewers for their valuable comments and advices.

References

[1] M. Landy and J. A. Movshon, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, vol. 1. Cambridge, MA, USA: MIT Press, 1991, pp. 3–20.

[2] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[3] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 43–54.

[4] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 31–42.

[5] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. CTSR 2005-02, 2005, pp. 1–11.

[6] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. IEEE Int. Conf. Comput. Photogr.*, Apr. 2009, pp. 1–8.

[7] Lytro, Mountain View, CA, USA. *The Lytro Camera*. Accessed: May 15, 2018. [Online]. Available: http://www.lytro.com/

[8] Raytrix, San Diego, CA, USA. *3D Light Field Camera Technology*. Accessed: May 15, 2018. [Online]. Available: http://www.raytrix.de/

[9] *Lytro Immerge*. Accessed: May 15, 2018. [Online]. Available: https://www.lytro.com/

[10] C. Zhang, G. Hou, Z. Sun, T. Tan, and Z. Zhou, "Light field photography for iris image acquisition," in *Proc. 8th Chin. Conf. Biometric Recognit. (CCBR)*. Jinan, China: Springer, 2013, pp. 345–352.

[11] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[12] C. Zhang, G. Hou, Z. Sun, and T. Tan, "Efficient auto-refocusing of iris images for light-field cameras," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep./Oct. 2014, pp. 1–7.

[13] R. Raghavendra, K. Raja, and C. Busch, "Exploring the usefulness of light field cameras for biometrics: An empirical study on face and iris recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 5, pp. 922–936, May 2016.

[14] T. G. A. Lumsdaine, "Full resolution lightfield rendering," Indiana Univ., Bloomington, IN, USA, Tech. Rep., Jan. 2008, pp. 1–12.

[15] T. E. Bishop, S. Zanetti, and P. Favaro, "Light field superresolution," in *Proc. IEEE Int. Conf. Comput. Photogr.*, Apr. 2009, pp. 1–9.

[16] T. G. Georgiev and A. Lumsdaine, "Superresolution with plenoptic 2.0 cameras," in *Frontiers Opt. Laser Sci. 25th/Fall OSA Opt. Photon. Tech. Dig.*, 2009, paper STuA6. [Online]. Available: www.tgeorgiev.net

[17] J. Lim, H. Ok, B. Park, J. Y. Kang, and S. Lee, "Improving the spatail resolution based on 4D light field data," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 1173–1176.

[18] F. P. Nava and J. P. Luke, "Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera," in *Proc. 3DTV Conf., True Vis.-Capture, Transmiss. Display 3D Video*, May 2009, pp. 1–4.

[19] S. Wanner and B. Goldluecke, "Spatial and angular variational super-resolution of 4D light fields," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 608–621.

[20] C.-K. Liang and R. Ramamoorthi, "A light transport framework for lenslet light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, Mar. 2015, Art. no. 16.

[21] H.-G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1547–1555.

[22] J. Li, M. Lu, and Z.-N. Li, "Continuous depth map reconstruction from light fields," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3257–3265, Nov. 2015.

[23] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 673–680.

[24] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2170–2181, Nov. 2016.

[25] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[26] F. Liu, G. Hou, Z. Sun, and T. Tan, "High quality depth map estimation of object surface from light-field images," *Neurocomputing*, vol. 252, pp. 3–16, Aug. 2017.

[27] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 121–138.

[28] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 57–65.

[29] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 848–852, Jun. 2017.

[30] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 235–243.

[31] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[32] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Comput. Vis.*, vol. 1, no. 1, pp. 7–55, 1987.

[33] R. Ng, "Fourier slice photography," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 735–744, 2005.

[34] F. P. Nava, J. G. Marichal-Hernández, and J. M. Rodríguez-Ramos, "The discrete focal stack transform," in *Proc. IEEE Eur. Signal Process. Conf.*, Aug. 2008, pp. 1–5.

[35] D. E. Jacobs, J. Baek, and M. Levoy, "Focal stack compositing for depth of field control," Stanford Comput. Graph. Lab., Stanford, CA, USA, Tech. Rep. 2012-1, 2012.

[36] F. Pérez, A. Pérez, M. Rodríguez, and E. Magdaleno, "Fourier slice super-resolution in plenoptic cameras," in *Proc. IEEE Int. Conf. Comput. Photography*, Apr. 2012, pp. 1–11.

[37] F. Pérez, A. Pérez, M. Rodríguez, and E. Magdaleno, "Super-resolved Fourier-slice refocusing in plenoptic cameras," *J. Math. Imag. Vis.*, vol. 52, no. 2, pp. 200–217, 2015.

[38] F. Pérez, A. Pérez, M. Rodríguez, and E. Magdaleno, "Lightfield recovery from its focal stack," *J. Math. Imag. Vis.*, vol. 56, no. 3, pp. 573–590, 2016.

[39] Y. Wang, G. Hou, Z. Sun, Z. Wang, and T. Tan, "A simple and robust super resolution method for light field images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2016, pp. 1459–1463.

[40] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, "Modeling the calibration pipeline of the Lytro camera for high quality light-field image reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3280–3287.

[41] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.*, vol. 32, no. 4, p. 46, 2013.

[42] R. A. Farrugia, C. Galea, and C. Guillemot, "Super resolution of light field images using linear subspace projection of patch-volumes," *IEEE J. Sel. Topics Quantum Electron.*, vol. 11, no. 7, pp. 1058–1071, Oct. 2017.

[43] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 8, pp. 824–831, Aug. 1994.

[44] M. Lee and Y. W. Tai, "Robust all-in-focus super-resolution for focal stack photography," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1887–1897, Apr. 2016.

[45] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. Annu. Workshop Vis., Model. Vis.*, 2013, pp. 1–8.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[47] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[48] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[49] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[51] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[52] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.

[53] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[54] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1637–1645.

[55] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1–9.

[56] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 624–632.

[57] S. Baker and T. Kanade, "Super-resolution optical flow," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 1999.

[58] D. Mitzel, T. Pock, T. Schoenemann, and D. Cremers, "Video super resolution using duality based TV-L1 optical flow," in *Proc. Joint Pattern Recognit. Symp.*, 2009, pp. 432–441.

[59] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[60] J. Caballero *et al.* (2016). "Real-time video super-resolution with spatio-temporal networks and motion compensation." [Online]. Available: https://arxiv.org/abs/1611.05250

[61] Theano Development Team *et al.* (2016). "Theano: A Python framework for fast computation of mathematical expressions." [Online]. Available: https://arxiv.org/abs/1605.02688

[62] K. Mitra and A. Veeraraghavan, "Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 22–28.

[63] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.

[64] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[65] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.

Yunlong Wang received the B.E. and M.S. degrees from the Department of Automation, University of Science and Technology of China, where he is currently pursuing the Ph.D. degree. He is currently with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China. His research focuses on pattern recognition, machine learning, light field photography, and biometrics.

Fei Liu received the B.Sc. and M.Sc. degrees in educational technology from Shandong Normal University in 2007 and 2010, respectively, and the Ph.D. degree from the University of Chinese Academy of Sciences, China. She is currently holding the post-doctoral position with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. Her research focuses on light-field imaging, biometrics, and 3D modeling.

Kunbo Zhang received the B.E. degree in automation from the Beijing Institute of Technology in 2006, and the M.Sc. and Ph.D. degrees in mechanical engineering from The State University of New York at Stony Brook, NY, USA, in 2008 and 2011, respectively. From 2011 to 2016, he was a Machine Vision Research and Development Engineer of the Advanced Manufacturing Engineering Group, Nexteer Automotive, Saginaw, MI, USA. He is currently holding a post-doctoral position with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China. His current research interests focus on light-field photography, biometric imaging, robot vision, human-robot interaction, and intelligent manufacturing.

Guangqi Hou received the B.E. degree in optoelectronics technology, the M.S. degree in physical electronics, and the Ph.D. degree in optical engineering from the Beijing Institute of Technology in 2001, 2004, and 2011, respectively. He joined the Institute of Automation, Chinese Academy of Sciences, China, in 2011, as a Post-Doctoral Researcher. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. His research focuses on smart imaging, high-resolution imaging, light-field photography, computational photography, and real-time image process system.

Zhenan Sun (M'07) received the B.E. degree in industrial automation from the Dalian University of Technology in 1999, the M.S. degree in system engineering from the Huazhong University of Science and Technology in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, China, in 2006. He is currently a Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. His research focuses on biometrics and pattern recognition. He is also an Associate Editor of IEEE TIFS of IEEE Biometrics Compendium. He is a Fellow of IAPR.

Tieniu Tan (F'04) received the B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the M.Sc. and Ph.D. degrees in electronic engineering from the Imperial College of Science, Technology and Medicine, London, U.K., in 1986 and 1989, respectively. He is currently a Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China. His current research interests include biometrics, image and video understanding, information hiding, and information forensics. He is a Fellow of IAPR.