Spatial-Angular Attention Network for Light Field Reconstruction

Gaochang Wu, Yebin Liu, *Member, IEEE,* Lu Fang, *Senior Member, IEEE,* and Tianyou Chai, *Fellow, IEEE*

***Abstract*—Learning-based light field reconstruction methods demand in constructing a large receptive field by deepening the network to capture correspondences between input views. In this paper, we propose a spatial-angular attention network to perceive correspondences in the light field non-locally, and reconstruct high angular resolution light field in an end-to-end manner. Motivated by the non-local attention mechanism [1], [2], a spatial-angular attention module specifically for the high-dimensional light field data is introduced to compute the responses from all the positions in the epipolar plane for each pixel in the light field, and generate an attention map that captures correspondences along the angular dimension. We then propose a multi-scale reconstruction structure to efficiently implement the non-local attention in the low spatial scale, while also preserving the high frequency components in the high spatial scales. Extensive experiments demonstrate the superior performance of the proposed spatial-angular attention network for reconstructing sparsely-sampled light fields with non-Lambertian effects.**

***Index Terms*—Light field reconstruction, deep learning, attention mechanism.**

I. Introduction

TROUGH capturing both intensities and directions from sampled light rays, light field achieves high-quality view synthesis without the need of complex and heterogeneous information (e.g., geometry and texture). More importantly, benefited from the light field rendering technology [3], light field is capable of producing photorealistic views in real-time, regardless of the scene complexity or non-Lambertian effect. This high quality rendering usually requires light fields with disparities between adjacent views to be less than one pixel, i.e., the so-called densely-sampled light field (DSLF). However, typical DSLF capture either suffers from a long period of acquisition time (e.g., DSLF gantry system [3]) or falls into the well-known resolution trade-off problem, i.e., the light fields are sampled sparsely in either the angular [4] or the spatial domain [5] due to the limitation of the sensor resolution [6].

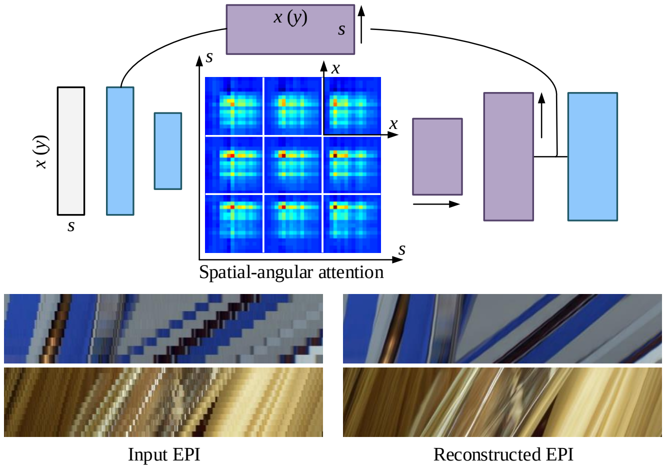


Fig. 1. We propose a spatial-angular attention module embedded in a multi-scale reconstruction structure for learning-based light field reconstruction. The network perceives correspondence pixels in a non-local manner, and is able to produce high quality reconstruction using sparse input. In the bottom results, the input light fields are upsampled by using nearest interpolation for better demonstration. Light fields courtesy of Moreschini *et al.* [15] and Adhikarla *et al.* [16].

Recently, a more promising way is the fast capturing of a sparsely-sampled (angular domain) light field followed by a direct reconstruction or a depth-based view synthesis [7], [8] with advanced deep learning techniques. On the one hand, typical learning-based reconstruction methods [9], [10], [11] employ multiple convolutional layers to map the low angular resolution light field to the DSLF. But due to the limited perceptive range of convolutional filters [12], the networks will fail to collect enough information among the correspondences when dealing with large disparities, leading to aliasing effects in the reconstructed light field. On the other hand, depth-based view synthesis methods address the large disparity problem through plane sweep (depth estimation), and then synthesize novel views using learning-based prediction [7], [13], [14]. However, such methods require depth consistency along the angular dimension, and thus, often fail to solve the depth ambiguity caused by the non-Lambertian effects.

In this paper, we propose a Spatial-Angular Attention Network, termed as SAA-Net, to achieve DSLF reconstruction from a sparse input. The proposed SAA-Net perceives correspondences in the Epipolar Plane Image (EPI) in a non-local manner, solving the aforementioned non-Lambertian effect and large disparity in a unified framework (Sec. IV). Specifically, the SAA-Net is composed by two parts, a spatial-angular attention module and a U-net backbone. Motivated by the

This work was supported by the Major Program of the NSFC grant No. 61991400, NSFC grant No.61991401, No.61827805, No.61531014, and Fundamental Research Funds for the Central Universities No. 100802004. (*Corresponding author: Tianyou Chai.*)

Gaochang Wu and Tianyou Chai are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, PR China (email: ahwgc2009@163.com; tychai@mail.neu.edu.cn).

Yebin Liu is with Department of Automation, Tsinghua University, Beijing 100084, PR China (email: liuyebin@mail.tsinghua.edu.cn).

Fang Lu is with Tsinghua-Berkeley Shenzhen Institute, Shenzhen 518055, P. R. China. (email: fanglu@sz.tsinghua.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes a video related to the work. Contact ahwgc2009@163.com for further questions about this work.

non-local attention mechanism in [1], [2], for each pixel in the input light field, the Spatial-Angular Attention Module (termed as SAAM for short) computes the responses from all the positions in its epipolar plane, and produces an attention map that records the correspondences along the angular dimension, as shown in Fig. 1 (top). This correspondence information in the attention map is then applied to guide the reconstruction in the angular dimension via multiplication and deconvolution.

To efficiently perform the non-local attention, we propose a convolutional neural network with multi-scale reconstruction structure. The network follows the basic architecture of the U-net, i.e., an encoder-decoder structure with skip connections. The encoder compresses the input light field in the spatial dimensions and removes redundancy information for the SAAM. Rather than simply reconstruct the light field at the end of the network, we propose a multi-scale reconstruction structure by performing deconvolution along the angular dimension in each skip connection branch, as shown in Fig. 1 (top). The proposed multi-scale reconstruction structure maintains the view consistency in the low spatial scale while preserving fine details in the high spatial scales.

For the network training, we propose a spatial-angular perceptual loss that is specifically designed for the high-dimensional light field data (Sec. V). Rather than computing the high-level feature loss [17], [18] by feeding each view in the light field to a 2D CNN (e.g., the commonly-used VGG [19]), we pretrain a 3D auto-encoder that considers the consistency in both the spatial and angular dimensions of the light field. We demonstrate the superiority of the SAA-Net by performing extensive evaluations on various light field datasets. The proposed network presents high-quality DSLF on challenge cases with both non-Lambertian effects and large disparities, as illustrated in Fig. 1 (bottom). In summary, we make the following contributions[1]:

- A spatial-angular attention module that perceives correspondences non-locally in the epipolar plane;
- A multi-scale reconstruction structure for efficiently performing the non-local attention in the low spatial scale while also preserving the high frequencies;
- A spatial-angular perceptual loss specifically designed for the high-dimensional light field data.

II. Related Work

A. Light Lield Reconstruction

First, we will give a brief review on researches of light field view synthesis (or view synthesis) depending on whether the depth information is explicitly used.

Depth image-based view synthesis. Typically, these kind of approaches first estimate the depth of a scene [20], [21], [22], [23], then warp and blend the input views to synthesize a novel view [13], [7], [24], [8]. Conventional light field depth estimation approaches follow the pipeline of stereo matching [25], i.e., cost computation, cost aggregation (or cost volume filtering) and post refinement. The main different is that light field converts the disparity from the discrete space into a continuous space [26], deriving various depth cues specifically for a light field, e.g., structure tensor-based local direction estimation [26], depth from defocus [20], [21]. Also, some learning-based approaches incorporate the depth estimation pipeline described above with 2D convolution-based feature extraction, 3D convolution-based cost volume refinement and depth regression [27], [28] For novel view synthesis, input views are warped to the novel viewpoints with sub-pixel accuracy using bilinear interpolation and blended in different manners, e.g., total variation optimization [26], soft blending [24] and learning-based synthesis [29].

Recently, researchers mainly focus on the studies for maximizing the quality of synthesized views based on the deep learning technique. Flynn *et al.* [13] proposed a learning-based method to synthesize novel views with predicted probabilities and colors for each depth plane. Kalantari *et al.* [7] further employed a sequential network to infer depth (disparity) and color, and optimized the model via end-to-end training. Shi *et al.* [8] developed a convolutional network that fuses low-level pixels and high-level features in a unified framework. Zhou *et al.* [30] introduced a learning-based MultiPlane Image (MPI) representation that infers a novel view by the alpha blending of different images. Mildenhall *et al.* [14] further proposed to use multiple MPIs to synthesize a local light field.

Reconstruction without explicit depth. These kind of approaches treat the problem of light field reconstruction as the approximation of plenoptic function. In the Fourier domain, the sparse sampling in the angular dimension produces overlaps between the original spectrum and its replicas, leading to aliasing effect. Classical approaches [31], [32] consider a reconstruction filter (usually in a wedge shape) to extract the original signal while filtering the aliasing high-frequency components. For instance, Vagharshakyan *et al.* [33] utilized an adapted discrete shearlet transform in the Fourier domain to remove the high-frequency spectra that introduce aliasing effects. Shi *et al.* [34] performed DSLF reconstruction as an optimization for sparsity in the continuous Fourier domain.

In recent years, some learning-based approaches were also proposed for depth-independent reconstruction [9], [10], [11], [35]. Zhu *et al.* [36] proposed an auto-encoder that combines convolutional layers and convLSTM layer [37]. For explicitly addressing the aliasing effects, Wu *et al.* [10] took advantage of the clear texture structure of the EPI and proposed a "blur-restoration-deblur" framework. However, when applying a large blur kernel for large disparities, the approach tends to fail at recovering the high-frequency details, and thus leading to the blur effect. Wang [35] further proposed to apply a 3D CNN that takes a 3D slice as the input. Yeung *et al.* [11] directly fed the entire 4D light field into a 4D convolutional network, and applied a coarse-to-fine model to iteratively refine the spatial and angular dimensions of the light field. Wu *et al.* [38] proposed an evaluation network for EPIs with different shear amount, termed as sheared EPI structure. In this structure, the depth information is implicitly used to select a well reconstructed EPI. However, the performance of the network is limited due to the finite perceptive field of the convolutional neurons, especially when handling the large disparity problem.

[1] We will release the source code of this work upon acceptance.

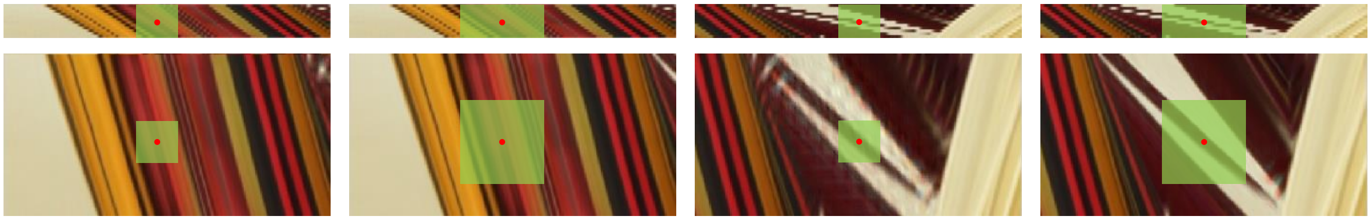


Fig. 2. Analysis of reconstruction quality in terms of the network receptive field and disparity range of the scene. For a scene with small disparities, both networks (a) with small receptive field (27 × 27 pixels) and (b) with large receptive field (53 × 53 pixels) are able to reconstruct high-quality light field. However, for a scene with large disparities, network with small receptive field suffers from severe aliasing effects, as shown in (c). While network with large receptive field can still produce plausible results, as shown in (d). We show the sparely-sampled inputs on the top row and the reconstructed on the bottom. The receptive field of each network is visualized with green box. The input EPIs are stretched along the angular dimension for better demonstration.

B. Attention Mechanism

Attention was first built to imitate the mechanism of human perception that mainly focuses on the salient part [39], [40], [41]. Vaswani *et al.* [42] indicated that the attention mechanism is able to solve the long term dependency problem even without being embedded in the backbone of a recurrent or C-NN. Therefore, the attention mechanism is recently developed to enabling the non-local perception in the spatial or temporal dimension [43].

To achieve the feature of non-local perception, Hu *et al.* [44] and Woo *et al.* [45] proposed to use a global pooling (max-pooling or average-pooling) followed by a multi-layer perceptron to aggregate the entire information in the spatial dimension. Tsai *et al.* [28] introduced an attention module in the angular dimension to weight the contribution of each view in a light field. Vaswani *et al.* [42] proposed to use a weighted average of the responses from all the positions with respect to a certain position in the latent space, which is called self-attention (also known as intra-attention). Alternatively, Wang *et al.* [1] achieved the self-attention by using matrix multiplication between reshaped feature maps, and is termed as non-local attention. For a high-dimensional task like video classification, the proposed module reshapes the 4D tensor (time, height, width and channel) into a 2D matrix. Zhang *et al.* [2] further extended this idea into a Generative Adversarial Network (GAN). Rather than using the non-local attention mechanism, Wang *et al.* [46] proposed a parallax attention module to compute the response across two stereo images. For each epipolar line in the stereo images (feature maps), the 2D matrices (width and channel) are multiplied to produce a sparse attention map that implies correspondences. In this paper, we extend the non-local attention mechanism to high dimensional light field data. For each pixel in the input 3D light field, the attention is computed in the 2D epipolar plane, rather than the epipolar line in [46] or the entire 3D data space in [1].

III. Problem Analysis

In the following analysis, we empirically show that the performance of a learning-based light field method is closely related to the perception range of its neuron (or convolutional filter), especially when addressing the large disparity problem. Deep neural network is proved to be a powerfull technique in solving ill-posed inverse problems [47]. In the light field reconstruction problem, the performance of a deep neural network mainly depends on two factors: disparity range of the scene (input light field) and network structure. Since the first factor is normally unalterable once the light field is acquired, typical deep learning-based approaches pursue a more appropriate architecture for higher performance [9], [7], [10], [11]. Among those deep learning-based approaches, the depth-based view synthesis methods convert the feature maps into a physically meaningful depth map, while depth-independent methods directly map them to novel views. Essentially, both of two kind of approaches employ convolutional filter to generate responses (feature maps) between correspondence pixels.

To quantify the measurement of the capability to capture correspondences, we apply the concept of receptive field introduced in [12], [48]. The receptive field measures the number of pixels that are connected to a particular filter in the CNN, i.e., the number of correspondence pixels perceived by the convolutional filter.

We analyse the reconstruction qualities of two networks with the same structure but different receptive field sizes, as illustrated in Fig. 2. For a scene with small disparity (about 3 pixels in the demonstrated example), both networks with small receptive field (27 × 27 pixels) and large receptive field (53 × 53 pixels) can reconstruct high angular resolution light fields (EPIs) with view consistency, as shown in Fig. 2(a) and Fig. 2(b). However, for a scene with large disparity (about 9 pixels), the network with small receptive field is not able to collect enough information from the corresponding pixels of its center point, as shown clearly at the top of Fig. 2(c). Note that the actual size of the receptive field can be smaller than its theoretical size [48]. Consequently, the reconstructed result suffers from severe aliasing effects, as shown at the bottom of Fig. 2(c). In comparison, the network with large receptive field can produce high quality result. In this example, the input EPIs are stretched along angular dimension for better demonstration.

Due to the limitation of parameter amount, it is intractable to expand the receptive field by stubbornly deepening the network or enlarging the filter size. The fundamental idea of the proposed approach is to design a light field reconstruction network that catches the correspondences non-locally across the spatial and angular dimensions of the light field via non-local perception. We achieve this with two aspects: 1) a spatial-

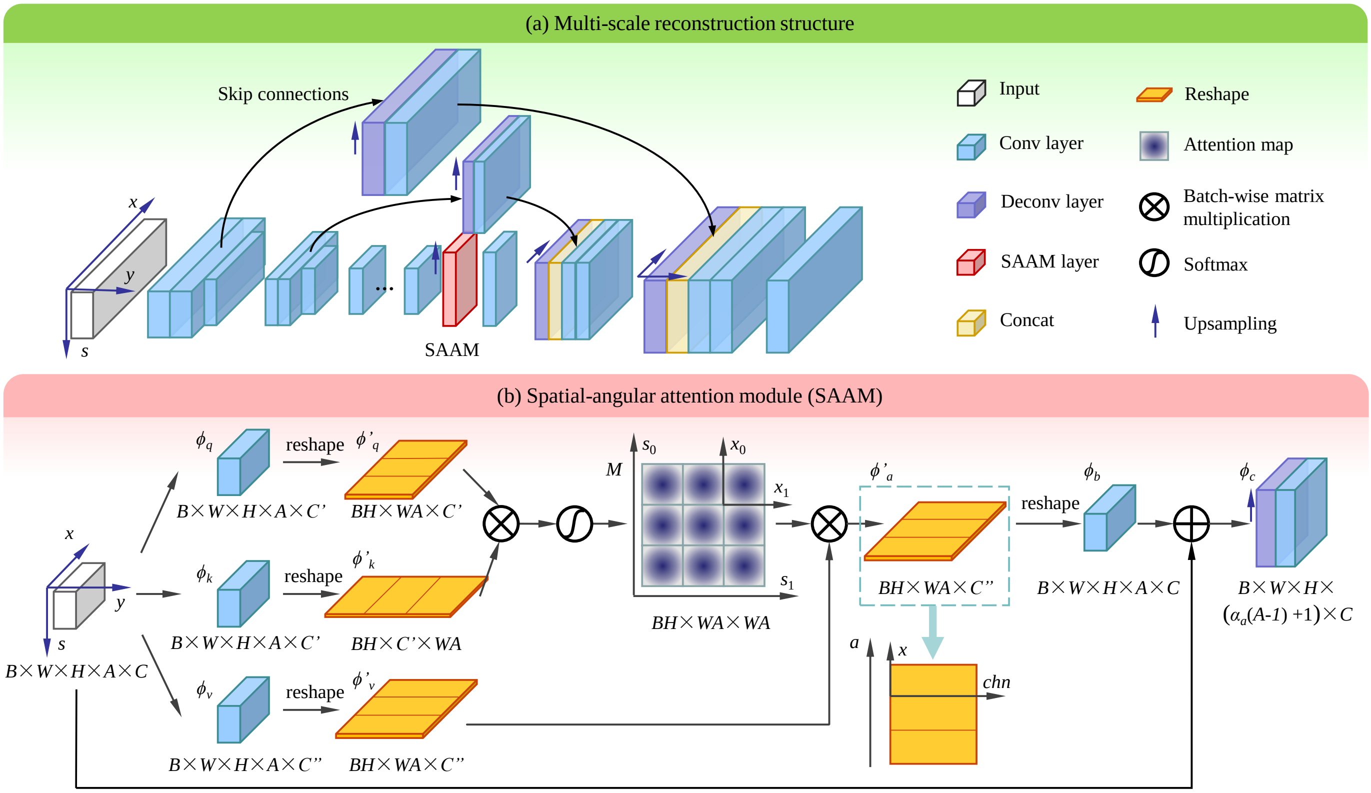


Fig. 3. Architecture of the proposed Spatial-Angular Attention Network (SAA-Net). The SAA-Net is composed of (a) a multi-scale reconstruction structure, and (b) a Spatial-Angular Attention Module (SAAM). The input is a 3D slice ($L(u, v, s)$ or $L(v, u, t)$) of the light field.The batch and channel dimensions are omitted in the figure.

angular attention module inspired by the non-local attention mechanism [1], [2]; and 2) an encoder-decoder network that can reduce the redundancies in the light field so that the non-local perception can be implemented efficiently.

IV. Spatial-Angular Attention Network

In this section, we first introduce the overall architecture of the proposed Spatial-Angular Attention Network for light field reconstruction, which is termed as SAA-Net. We then present the proposed spatial-angular attention module that is specifically designed for disentangling the disparity information with a non-local perception. The input of the SAA-Net is a 3D light field slice with two spatial dimensions and one angular dimension, i.e., $L(x, y, s)$ (or $L(y, x, t)$). By splitting light field into 3D slices, the proposed network can be adopted for not only 3D light fields from a single-degree-of-freedom gantry system but also 4D light fields from plenoptic camera and camera array system.

For a 4D light field $L(x, y, s, t)$, we adopt a hierarchical reconstruction strategy similar with that in [10]. The strategy first reconstruct 3D light fields using slices $L_{t^*}(x, y, s)$ and $L_{s^*}(y, x, t)$, then use the 3D light fields from the synthesized views to generate the final 4D light field.

A. Network Architecture

We propose a multi-scale reconstruction structure to maintain the view consistency (i.e., continuity in the angular dimension) in the low spatial scale while preserving fine details in the high spatial scales. The backbone of the proposed SAA-Net follows the encoder-decoder structure with skip connections, also known as U-net, as shown in Fig. 3(a). But the proposed SAA-Net has two particular differences: 1) We use deconvolution along the angular dimension in each skip connection branch before the concatenation in the decoder part; 2) We apply convolution layers with stride specifically in the spatial dimensions of the light field. Table I provides the detailed configuration of the proposed SAA-Net.

The **encoder** part of the SAA-Net construct multi-scale light field features and reduces the redundant information in the spatial dimension to alleviate the computational and GPU memory costs for the non-local perception in the spatial-angular attention module. We use two convolutional layers (3D) with stride $[2, 2]$ and $[2, 1]$ to downsample the spatial resolution of the light field with ratio 4 and 2 along the width and height dimension, respectively. Before each downsampling, two 3D convolutional layers with filter sizes $3 \times 1 \times 3$ and $1 \times 3 \times 3$ (width, height and angular) are employed to take place of a single convolutional layer with filter size $3 \times 3 \times 3$, reducing $1/3$ parameters without performance degradation.

The **skip connections** are fed with the feature layers before the downsampling in each encoder level, as shown in Fig. 3(a), which have full and half spatial resolutions. For each skip connection, a deconvolution layer (also known as transpose convolution layer) is then applied to upsample the feature map in the angular dimension, followed by a $1 \times 1 \times 1$ convolution. For the reconstruction of a 3D light field $L(x, y, s)$, the angular

TABLE I
DETAIL CONFIGURATION OF THE PROPOSED SAA-NET, WHERE k DENOTES THE KERNEL SIZE, s THE STRIDE, chn THE NUMBER OF CHANNELS, CONV THE 3D CONVOLUTION LAYER, DECONV THE 3D DECONVOLUTION LAYER AND CONCAT THE CONCATENATION.

Layer	k	s	chn	Input
Encoder				
Conv1_1	$3 \times 1 \times 3$	-	1/24	$L(x, y, s)$
Conv1_2	$1 \times 3 \times 3$	-	24/24	Conv1_1
Conv1_3	$3 \times 3 \times 1$	$[2, 2, 1]$	24/48	Conv1_2
Conv2_1	$3 \times 1 \times 3$	-	48/48	Conv1_3
Conv2_2	$1 \times 3 \times 3$	-	48/48	Conv2_1
Conv2_3	$3 \times 1 \times 1$	$[2, 1, 1]$	48/96	Conv2_2
Conv3_1	$1 \times 1 \times 1$	-	96/48	Conv2_3
Conv3_2	$3 \times 1 \times 3$	-	48/48	Conv3_1
Conv3_3	$1 \times 3 \times 3$	-	48/48	Conv3_2
Conv3_4	$3 \times 1 \times 3$	-	48/48	Conv3_3
Conv3_5	$1 \times 3 \times 3$	-	48/48	Conv3_4
Skip connection				
Deconv4_1	$3 \times 1 \times 7$	$[1, 1, \alpha_a]$	24/24	Conv1_2
Conv4_2	$1 \times 1 \times 1$	-	24/24	Deconv4_1
Deconv5_1	$3 \times 1 \times 7$	$[1, 1, \alpha_a]$	48/48	Conv2_2
Conv5_2	$1 \times 1 \times 1$	-	48/48	Deconv5_1
SAAM				
Decoder				
Conv6_1	$1 \times 1 \times 1$	-	48/96	SAAM
Deconv6_2	$4 \times 1 \times 1$	$[2, 1, 1]$	96/48	Conv6_1
Concat1	-	-	-	Conv6_1;Conv4_2
Conv6_3	$3 \times 1 \times 3$	-	48/48	Concat1
Conv6_4	$1 \times 3 \times 3$	-	48/48	Conv6_3
Deconv7_1	$4 \times 4 \times 1$	$[2, 2, 1]$	48/24	Conv6_4
Concat2	-	-	-	Conv7_1;Conv5_2
Conv7_2	$3 \times 1 \times 3$	-	24/24	Concat2
Conv7_3	$1 \times 3 \times 3$	-	24/24	Conv7_2
Conv8	$3 \times 3 \times 3$	-	24/1	Conv7_3

information can be mainly extracted from the 2D EPI $E(x, s)$, therefore, the filter size in each deconvolution layer is set to $3 \times 1 \times 7$.

The **decoder** part of the SAA-Net upsamples the feature map from the spatial-angular attention module by using two deconvolution layer with stride $[2, 1]$ and $[2, 2]$ in the spatial dimensions (width and height). The decoder also receives information from the skip connections by concatenating the them along the channel dimension [49], as shown in Fig. 3(a). We then use two 3D convolutional layers with filter sizes $3 \times 1 \times 3$ and $1 \times 3 \times 3$ to compress the channel numbers in each level of the decoder. This can be considered as the blending of the light field features from different reconstruction scale. Note that all the reconstructions (upsampling operations) in the angular dimension are performed by the skip connections or the spatial-angular attention module, where latter will be introduced in the following subsection.

B. Spatial-Angular Attention Module

Inspired by the non-local attention mechanism in [1], [2], we propose a Spatial-Angular Attention Module (SAAM) to disentangling the disparity information in light field. The main differences between the proposed SAAM and the previous non-local attention [1], [2] are as follows: 1) Since the disparity information is encoded in the EPI, the non-local attention mechanism is performed in the 2D epipolar plane rather than the entire 3D space; 2) Taking advantage of the non-local perception of the EPI, we embed light field reconstruction in the spatial-angular attention module.

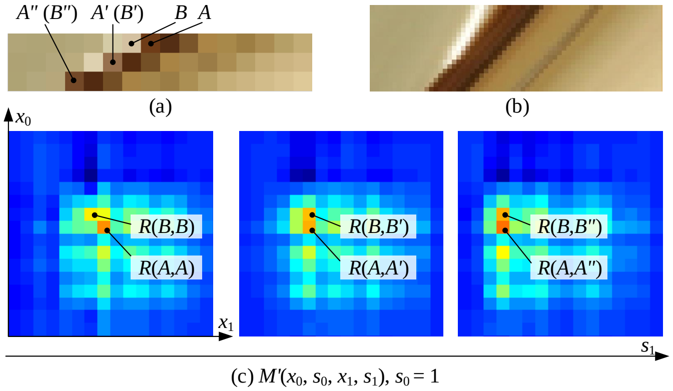


Fig. 4. Visualization of the attention map. (a) An EPI with a foreground point A and a background point B; (b) The corresponding high spatial-angular resolution EPI; (c) Three sub-maps extracted from the attention map. A point will have a strong response at the location of its correspondence in the attention map.

A straightforward choice of performing spatial-angular attention is to embed the attention module in each resolution scale of the U-net. However, implementing non-local perception in the full resolution light field (feature map) is intractable in terms of computation complexity and GPU memory. Alternatively, we insert the proposed SAAM between the encoder and decoder as shown in Fig. 3(b).

In a 3D convolutional network, the feature layer will be a 5D tensor $\phi \in \mathbb{R}^{B \times W \times H \times A \times C}$ (i.e., batch, width, hight, angular and channel). We first apply two convolution layers with kernel size $1 \times 1 \times 1$ to produce two feature layers ϕ_q and ϕ_k with size of $B \times W \times H \times A \times C'$. The channel number C' is set to be $\frac{C}{8}$ (i.e., $C' = 6$) for computation efficiency. Then the feature layers ϕ_q and ϕ_k are reshaped into 3D tensors ϕ'_q and ϕ'_k of shapes $BH \times WA \times C'$ and $BH \times C' \times WA$, respectively. In this way, we merge the angular and width dimensions (s and x or t and y in a light field) together to implement the non-local perception in the epipolar plane.

We apply batch-wise matrix multiplication between ϕ'_q and ϕ'_k and a softmax function to produce a attention map M as illustrated in Fig. 3(b). The attention map is composed of BH matrices with shape $WA \times WA$. Each matrix can be considered as a 2D expansion map of a 4D tensor $M' \in \mathbb{R}^{W \times A \times W \times A}$ (the batch and height dimensions are neglected). The point $M'(x_0, s_0, x_1, s_1)$ indicates the response of light field position $L(x_0, y, s_0)$ to position $L(x_1, y, s_1)$ in the latent space. In other words, the attention map is able to capture correspondence among all the views in the input 3D light field.

We demonstrate the non-local perception of the proposed SAAM by visualizing a part of the attention map as shown in Fig. 4. In this example, there are two points A and B with remarkable visual features as shown in Fig. 4(a). And their corresponding points in other views are marked as A' (A'') and B' (B''). As the viewpoint changes along the angular dimension, the background point B will be occluded by the foreground point A, which is demonstrated more obviously in Fig. 4(b). Fig. 4(c) shows three sub-maps extracted from

TABLE II
DETAIL CONFIGURATION OF THE PROPOSED SPATIAL-ANGULAR ATTENTION MODULE (SAAM), WHERE MATMUL DENOTES THE MATRIX MULTIPLICATION AND ADD THE ELEMENT-WISE ADDITION.

Layer	k	s	chn	Input
Conv1	$1 \times 1 \times 1$	-	48/6	Encoder
Conv2	$1 \times 1 \times 1$	-	48/6	Encoder
Conv3	$1 \times 1 \times 1$	-	48/6	Encoder
Reshape1	-	-	-	Conv1
Reshape2	-	-	-	Conv2
Reshape3	-	-	-	Conv3
MatMul1	-	-	-	Reshape1;Reshape2
Softmax	-	-	-	MatMul1
MatMul2	-	-	-	Softmax;Reshape3
Reshape4	-	-	-	MatMul2
Conv4	$1 \times 1 \times 1$	-	24/48	Reshape4
Add	-	-	48/48	Encoder;Conv4
Deconv	$3 \times 1 \times 7$	$[1, 1, \alpha_a]$	48/48	Add
Conv_6	$1 \times 1 \times 1$	-	48/48	Deonv

the attention map M' with $s_0 = 1$ and $s_1 = 1, 2$ and 3, respectively. It can be clearly seen that a point will have the strongest response at the location of its correspondence in the attention map. For instance, the response $R(A, A')$ at the location $M'(8, 1, 6, 2)$ for the corresponding patch (A, A') (the middle sub-figure of Fig. 4(c)), and the response $R(A, A'')$ at the location $M'(8, 1, 4, 3)$ for the corresponding patch (A, A'') (the right sub-figure of Fig. 4(c)). For the occluded point B, the location of the maximum response changes from $M'(7, 1, 7, 1)$ to $M'(7, 1, 4, 3)$[2]. In this case, the attention module is able to locate the occluded point B'' through the surrounding pixels.

Similar with ϕ'_q and ϕ'_k, ϕ'_v is obtained by another $1 \times 1 \times 1$ convolution using input tensor ϕ, followed by the reshape operation. The main difference is that the channel number of the feature layer is $C'' = \frac{C}{2}$, i.e., $C'' = 24$ in our implementation. Another batch-wise matrix multiplication is applied between the attention map M and ϕ'_v, resulting a 3D tensor $\phi'_a \in \mathbb{R}^{BH \times WA \times C''}$. We then reshape ϕ'_a into a 5D tensor $\phi_a \in \mathbb{R}^{B \times W \times H \times A \times C''}$ and adopt a $1 \times 1 \times 1$ convolution to expand the channel dimension from C'' to C, generating a 5D tensor (or feature layer) $\phi_b \in \mathbb{R}^{B \times W \times H \times A \times C}$. We further multiply the feature layer ϕ_b by a trainable scale parameter (initialized as 0) and add back the input feature layer.

We implicitly adopt the non-local similarities or correspondences captured by the spatial-angular attention in the latent space and perform light field reconstruction by using deconvolution in the angular dimension, as shown in Fig. 3. The output of the SAAM is a 5D tensor $\phi_c \in \mathbb{R}^{B \times W \times H \times (\alpha_a(A-1)+1) \times C}$. By combining the proposed SAAM with the feature maps in the skip connections, the network is able to reconstruct light field with view consistency while also preserving the high frequency components. Detailed parameter setting of SAAM is listed in Table II.

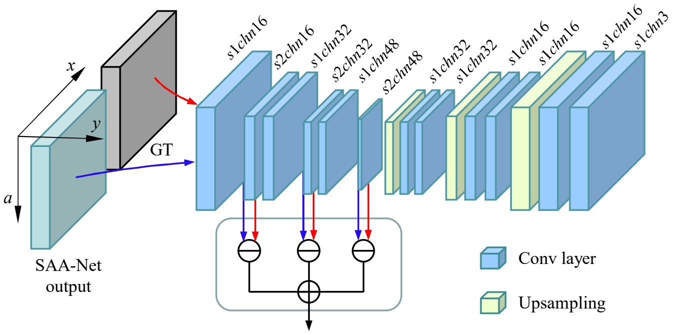


Fig. 5. Architecture of the 3D encoder-decoder network designed for the proposed spatial-angular perceptual loss.

V. NETWORK TRAINING

A. Spatial-Angular Perceptual Loss

Typical learning-based light field reconstruction or view synthesis methods optimize the network parameters by formulating a pixel-wise loss between the inferred image and the desired view (or EPI [10]). Recently, researches [30], [50], [51], [14] show that formulating the loss function in the high-level feature space will motivate the restoration of high-frequency components. This high-level feature loss, also known as perceptual loss, can be computed from part of the feature layers in the autologous network [17] or other pre-trained networks [18], such as the commonly-used VGG network [19].

In this paper, we propose a spatial-angular perceptual loss that is specifically designed for the high-dimensional light field data. Existing approaches [14], [51] for light field reconstruction apply perceptual loss between 2D sub-aperture images, neglecting the view consistency constraint in the angular dimension. Alternatively, we propose to use a 3D light field encoder to map the 3D light fields into high-dimensional feature tensors (width, height, angular and feature channel). To achieve this, we design another 3D encoder-decoder network (auto-encoder)[3] optimized by using unsupervised learning, i.e., the network is trained by inferring (compress and restore) the input light field. We then employ the encoder part to extract the high-level feature for the proposed spatial-angular perceptual loss. Note that the auto-encoder can be also generalized to 4D form. But given that some light field datasets have only one angular dimension (e.g., light fields from gantry system in [15]) and the proposed SAA-Net also takes 3D light field as input, we only adopts 3D convolution in the encoder and decoder.

Fig. 5 demonstrates the computation process of the proposed spatial-angular perceptual loss as well as the designed auto-encoder. We use 3D convolutional layers with kernel size $3 \times 3 \times 3$ to encode the 3D light field (the SAA-Net output or the ground truth) into latent representations. The encoder applies stride convolutional layers with stride 2 in each dimension to compress the light field from low-level pixel space into high-level feature space. The decoder employs bilinear upsampling

[2] Due to the subpixel disparity of point B, the actual location of the maximum response could be $M'(7, 1, 5.5, 2)$ in the middle sub-figure of Fig. 4(c).

[3] The architecture of the 3D auto-encoder for the perceptual loss is different with that of the SAA-Net.

and convolutional layers to restore the light field from the latent representations. Detailed configuration of each layer is shown in Fig. 5. The loss function $\mathcal{L}_{AE}$ for optimizing the auto-encoder is as

$$\mathcal{L}_{AE}(L_{HR}) = \|f_{AE}(L_{HR}) - L_{HR}\|_1,$$

where f_{AE} denotes the auto-encoder.

With the unsupervised learning, the encoder part is trained to extract high-frequency features in different scales. In this paper, we use the second, fourth and sixth layers in the encoder to form the spatial-angular perceptual loss

$$\mathcal{L}_{feat}(\hat{L}_{HR}, L_{HR}) = \sum_{l=2,4,6} \lambda_{feat}^{(l)} \|\phi_{ae}^{(l)}(\hat{L}_{HR}) - \phi_{ae}^{(l)}(L_{HR})\|_1,$$

where $\phi_{ae}^{(l)}(\cdot)$ $(l = 2,4,6)$ denotes the feature layers in the encoder, and $\lambda_{feat} = 0.2, 0.2, 0.1$ is a set of hyperparameters for the proposed spatial-angular perceptual loss, $\hat{L}_{HR}$ is the light field reconstructed by the SAA-Net and L_{HR} is the desired high-angular resolution light field.

To prevent the potential possibility that different light field patches are mapped to the same feature vector [17], our loss function also contains a pixel-wise term $\mathcal{L}_{pix}$ using Mean Absolute Error (MAE) between $\hat{L}_{HR}$ and L_{HR}, i.e.,

$$\mathcal{L}_{pix}(\hat{L}_{HR}, L_{HR}) = \|\hat{L}_{HR} - L_{HR}\|_1.$$

Then the final loss function $\mathcal{L}_{SAA}$ for training the SAA-Net is defined as

$$\mathcal{L}_{SAA} = \mathcal{L}_{pix} + \mathcal{L}_{feat}. \quad (1)$$

The two terms are weighted by the set of hyperparameters λ_{feat} in the perceptual loss.

B. Training Data

We use light fields from the Stanford (New) Light Field Archive [52] as the training dataset, which contains 12 light fields[4] with 17 × 17 views. Since the network input is 3D light fields, we can extract 17 $L(x,y,s)$ and 17 $L(y,x,t)$ in each 4D light field set. Similar with the data augmentation strategy proposed in [36], we augment the extracted 3D light fields using shearing operation [53]

$$L_d(x,y,s) = L(x + (s - \frac{S}{2}) \cdot d, y, s),$$

where S is the angular resolution of the 3D light field $L(x,y,s)$ and $L_d(x,y,s)$ is the resulting 3D light field with shear amount d. $L_d(y,x,t)$ can be obtained similarly. In practice, we use two shear amounts $d \pm 2$. The shearing-based data augmentation increases the number of training examples by 2 times. More importantly, the disparity effects in the augmented light field will be more obvious as shown in Fig. 6, enabling the network to address the large disparity problem.

To accelerate the training procedure and insure the same resolution between the input examples in the meantime, the extracted 3D light fields are cropped into sub-light fields of spatial resolution 64×24 (width and height for $L(x,y,s)$ or height and width for $L(y,x,t)$) with stride 40 pixels. About 6.7×10^5 examples can be extracted from the 3D light fields (original and augmented).

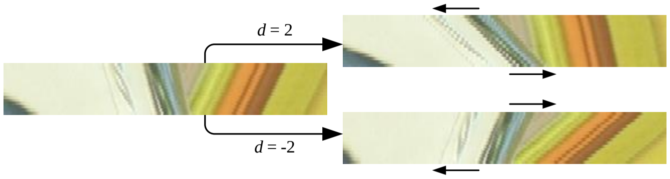


Fig. 6. An illustration of training data augmentation using shearing operation. For clear display, one of the spatial dimension in the 3D light field is ignored.

C. Implementation Details

Two models with reconstruction factors (upsampling scale in the angular dimension) $\alpha_a = 3, 4$ are trained. The input/output angular resolution of the training samples for these two models are 5/17 and 6/16, respectively. Although the reconstruction factor of the network is fixed, we can achieve a flexible upsampling rate through network cascade. The training is performed on the Y channel (i.e., the luminance channel) of the YCbCr color space. We initialize the weights of both convolution and deconvolution layers by drawing randomly from a Gaussian distribution with a zero mean and standard deviation 1×10^{-3}, and the biases by zero. The network is optimized by using ADAM solver [54] with learning rate of 1×10^{-4} ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and mini-batch size of 28. The training model is implemented using the *Tensorflow* framework [55]. The network converges after 8×10^5 steps of backpropagation, taking about 35 hours on a NVIDIA Quadro GV100.

VI. Evaluations

In this section, we evaluate the proposed SAA-Net on several datasets, including light fields from gantry system, light fields from plenoptic camera (Lytro Illum [6]). We mainly compare our approach with three state-of-the-arts learning-based methods by Kalantari *et al.* [7] (depth-based), Wu *et al.* [10] (without explicit depth) and Yeung *et al.* [11] (without explicit depth). To empirically validate the effectiveness of the proposed schemes, we perform ablation studies of our approach by training our network without the SAAM, without the multi-scale reconstruction structure and without the spatial-angular perceptual loss, respectively. The quantitative evaluations is reported by measuring the average PSNR and SSIM [56] values over the synthesized views of the luminance channel. For more quantitative and qualitative evaluations, please see the submitted video.

A. Evaluations on Light Fields from Gantry Systems

A gantry system capture a light field by mounting a conventional camera on a mechanical gantry. Typical gantry system takes minutes to hours (depending on the angular density) to take a light field. With a high quality DSLF reconstruction / view synthesis approach, the acquisition time

[4] The light field *Lego Gantry Self Portrait* is excluded from the training dataset since the moving camera may influence the reconstruction performance.

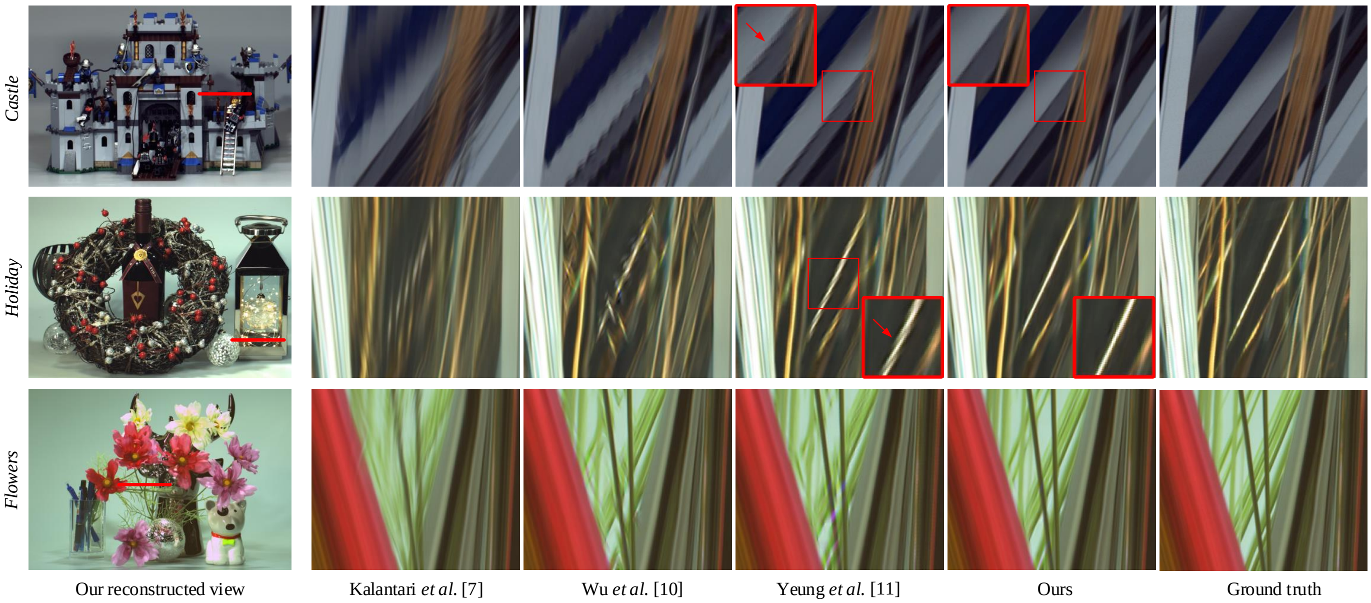


Fig. 7. Comparison of the results on the light fields from the CIVIT Dataset [15] (16× upsampling). The results show one of our reconstructed view, EPIs extracted from light fields reconstructed by each method.

TABLE III
QUANTITATIVE RESULTS (PSNR/SSIM) OF RECONSTRUCTED LIGHT FIELDS ON THE LIGHT FIELDS FROM THE CIVIT DATASET [15].

	Scale	*Seal & Balls*	*Castle*	*Holiday*	*Dragon*	*Flowers*	Average
Kalantari *et al.* [7]	8×	46.83 / 0.990	39.14 / 0.973	36.03 / 0.979	43.97 / 0.989	39.00 / 0.989	40.99 / 0.984
Wu *et al.* [10]		49.01 / 0.997	37.67 / 0.984	40.46 / 0.995	48.38 / 0.997	45.85 / 0.998	44.27 / 0.994
Yeung *et al.* [11]		49.83 / 0.997	40.84 / 0.993	41.16 / 0.996	48.61 / 0.997	47.83 / 0.997	45.65 / 0.996
Our proposed		**51.05 / 0.998**	**43.15 / 0.994**	**42.27 / 0.997**	**49.68 / 0.998**	**48.35 / 0.998**	**46.90 / 0.997**
Kalantari *et al.* [7]	16×	43.13 / 0.985	36.03 / 0.965	32.44 / 0.961	39.50 / 0.985	35.21 / 0.973	37.26 / 0.974
Wu *et al.* [10]		45.21 / 0.994	35.20 / 0.977	35.58 / 0.987	46.39 / 0.997	41.60 / 0.995	40.80 / 0.990
Yeung *et al.* [11]		44.38 / 0.992	37.86 / 0.989	36.06 / 0.988	45.52 / 0.997	42.30 / 0.994	41.22 / 0.992
w/o SAAM		46.85 / 0.995	37.78 / 0.989	36.17 / 0.988	47.10 / 0.998	42.98 / 0.996	42.18 / 0.993
w/o MSR structure		46.53 / 0.995	38.33 / 0.990	36.94 / 0.989	46.92 / 0.997	43.01 / 0.996	42.35 / 0.993
w/o SAP loss		49.02 / 0.996	40.69 / 0.992	38.97 / 0.992	48.23 / 0.997	44.46 / 0.997	44.27 / 0.995
Our proposed		**49.35 / 0.997**	**40.85 / 0.992**	**39.01 / 0.993**	**48.54 / 0.998**	**44.67 / 0.997**	**44.48 / 0.995**

will be considerably reduced. In this experiment, we use light fields from the CIVIT Dataset [15] (1×193 views of resolution 1280×720) and the MPI Light Field Archive [16] (1×101 views of resolution 960×720) with upsampling scales 8× and 16×. In this experiment, the performances in terms of both angular sparsity and non-Lambertian are taken into consideration. Since the vanilla version of the network by Yeung et al. [11] was specifically designed for 4D light fields, we modify the convolutional layers for the 3D input while remain the network architecture unchanged. The networks by Kalantari *et al.* [7] and Yeung et al. [11] are re-trained using the same training dataset as the proposed network. In the modified implementation, every 8 (6) views are applied to reconstruct (synthesize) a 3D light field of 22 (21) views for the networks of reconstruction factor $\alpha_a = 3$ ($\alpha_a = 4$). We use network cascade to achieve different upsampling scales, e.g., two cascades for 16× upsampling using a network of reconstruction factor $\alpha_a = 4$.

Fig. 7 shows the reconstruction results on three light fields, *Castle*, *Holiday* and *Flowers*, from the CIVIT Dataset [15] with upsampling scale 16× (disparity range $d_{\min} - d_{\max} =$ 14px). The first case and the third case have thin structures with complex occlusions. The depth and learning-based approach by Kalantari *et al.* [7] fail to estimate depth maps accurately enough to warp the input images, and the color CNN is not able to correct the misaligned views, producing ghosting artifacts as shown in the figure. For the second case, we demonstrate reconstructed EPIs in a highly non-Lambertian region, as shown in the figure. Caused by the depth ambiguity, the approach by Kalantari *et al.* [7] produces choppiness artifacts along the angular dimension. Due to the limited receptive field of the networks, the results by Wu *et al.* [10] and Yeung et al. [11] show aliasing effects in various degrees. Table III lists the quantitative measurements on the light fields from the CIVIT Dataset [15] with upsampling scale 8× and 16×.

Fig. 8 shows the reconstruction results on three light fields, *Bikes*, *FairyCollection* and *WorkShop*, from the MPI Light Field Archive [16] with upsampling scale 16× (disparity range up to 33.5px). The first case has complex structure occluded by the foreground bikes as shown in the top row of Fig. 8. The baseline methods fail to reconstruct the complex structure in the background. Among them, the depth and learning-based approach [7] fail to estimate a proper occlusion relation between the bikes and the background. The second scene is a non-Lambertian case, i.e., a refractive glass before the toys.

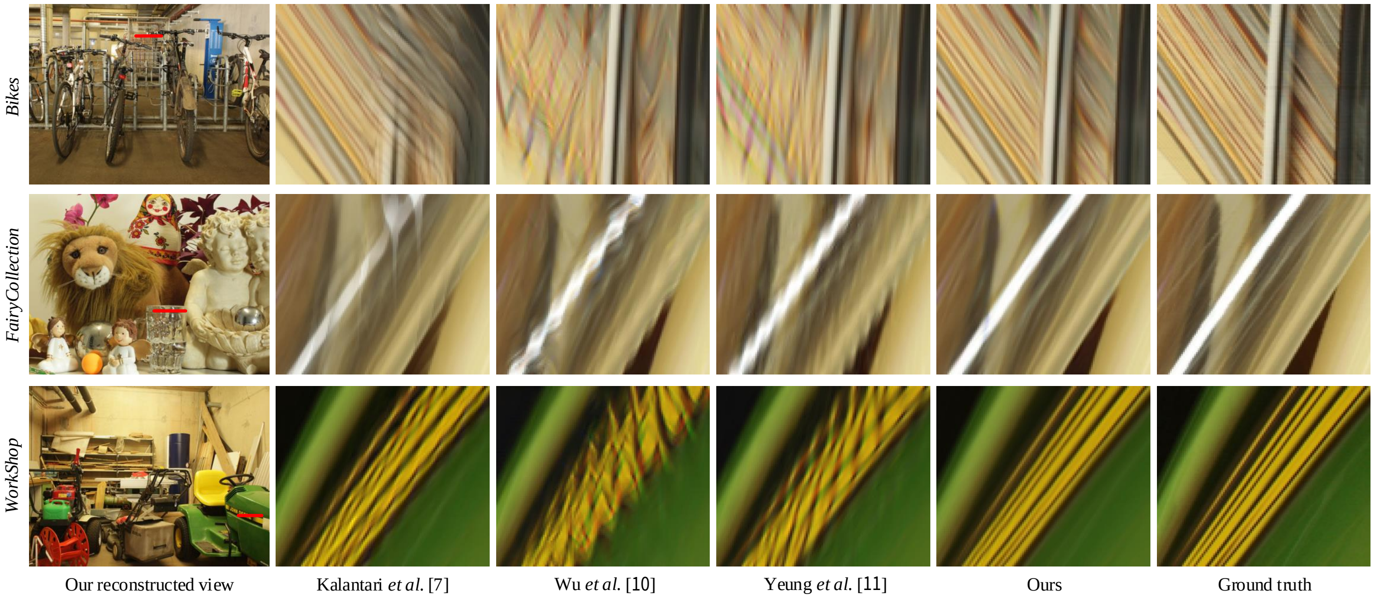


Fig. 8. Comparison of the results on the light fields from the MPI Light Field Archive [16] (16× upsampling).

TABLE IV
QUANTITATIVE RESULTS (PSNR/SSIM) OF RECONSTRUCTED LIGHT FIELDS ON THE LIGHT FIELDS FROM THE MPI LIGHT FIELD ARCHIVE [16].

	Scale	*Bikes*	*FairyCollection*	*LivingRoom*	*Mannequin*	*WorkShop*	Average
Kalantari *et al.* [7]	8×	34.83 / 0.969	36.66 / 0.977	46.35 / 0.991	40.62 / 0.983	38.66 / 0.986	39.42 / 0.981
Wu *et al.* [10]		38.39 / 0.990	40.32 / 0.992	45.48 / 0.996	43.26 / 0.995	41.55 / 0.995	41.80 / 0.994
Yeung *et al.* [11]		39.55 / 0.993	40.25 / 0.993	47.32 / 0.997	44.49 / 0.996	43.17 / 0.996	42.96 / 0.995
Our proposed		**40.53 / 0.995**	**42.23 / 0.995**	**47.96 / 0.997**	**45.02 / 0.996**	**45.29 / 0.997**	**44.21 / 0.996**
Kalantari *et al.* [7]	16×	30.67 / 0.935	32.39 / 0.952	41.62 / 0.973	37.15 / 0.970	33.94 / 0.971	35.15 / 0.960
Wu *et al.* [10]		31.22 / 0.951	30.33 / 0.942	42.43 / 0.991	39.53 / 0.989	33.49 / 0.977	35.40 / 0.970
Yeung *et al.* [11]		32.67 / 0.967	31.82 / 0.969	43.54 / 0.993	40.82 / 0.992	37.21 / 0.988	37.21 / 0.982
Our proposed		**36.01 / 0.985**	**36.13 / 0.982**	**46.45 / 0.997**	**41.08 / 0.993**	**39.11 / 0.992**	**39.76 / 0.990**

The approach by Kalantari *et al.* [7] cannot reconstruct the refractive object. And the reconstructed EPIs by the baseline methods [10], [11] appear severe aliasing effects. Table IV lists the quantitative measurements on the light fields from the MPI Light Field Archive [16] with upsampling scale 8× and 16×.

Ablation studies. We empirically validate the proposed approach by performing the following ablation studies. First, we replace the proposed SAAM with a simple transpose convolution layer, denoted as “w/o SAAM” for short. As show by the quantitative result in Table III, the average PSNR value decreases about 2.3dB without the SAAM. In the second ablation study, we use a typical 3D U-net as the backbone and remove the transpose convolution layer in the SAAM, denoted as “w/o MSR structure” for short (without the Multi-Scale Reconstruction structure). The angular reconstruction is simply realized by using transpose convolution at the end of the network. The performance of the network decreases about 2.1dB in terms of PSNR. In the last ablation study, we train the proposed SAA-Net simply by using the pixel-wise term (MAE loss) without the proposed spatial-angular perceptual loss, denoted as “w/o SAP loss” for short. The performance (PSNR) decreases about 0.21dB.

B. Evaluations on Light Fields from Lytro Illum

We evaluate the proposed approach using three Lytro light field datasets (113 light fields in total), the *30 Scenes* dataset by Kalantari *et al.* [7], and the *Reflective* and *Occlusions* categories from the Stanford Lytro Light Field Archive [57]. In this experiment, we reconstruct a 7×7 light field from 3×3 views (3× upsampling) and a 8×8 light field from 2×2 views (7× upsampling). Since the vanilla versions of the networks by Kalantari *et al.* [7], Yeung et al. [11], Wang *et al.* [35] and Meng *et al.* [51] are trained on Lytro light fields, we use their original parameters without re-training. Note that the proposed network is not fine-tuned on any Lytro light field datasets, and the results are produced by the same set of network parameters for both upsampling scales 3× and 7×.

We demonstrate two cases with relatively large disparities (maximum disparity up to 13px), *IMG1743* from the *30 Scenes* [7] and *Occlusions 23* from the *Occlusions* category [57], as shown in Fig. 9. In both cases, the reconstruction results by Wu *et al.* [10] and Yeung et al. [11] show ghosting artifacts around the region with large disparity (background in the *IMG1743* case, and foreground in the *Occlusions 23* case), which we believe are caused by the limited receptive field of their networks. The depth and learning-based approach by Kalantari *et al.* [7] produces plausible result in the first case, but appears tearing artifacts near the occlusion boundary as marked by the red arrow in the EPI. In the second case, the approach by Kalantari *et al.* [7] fail to estimate proper depth information, introducing misalignment as shown by the EPI. In comparison, the proposed SAA-Net provides reconstructed

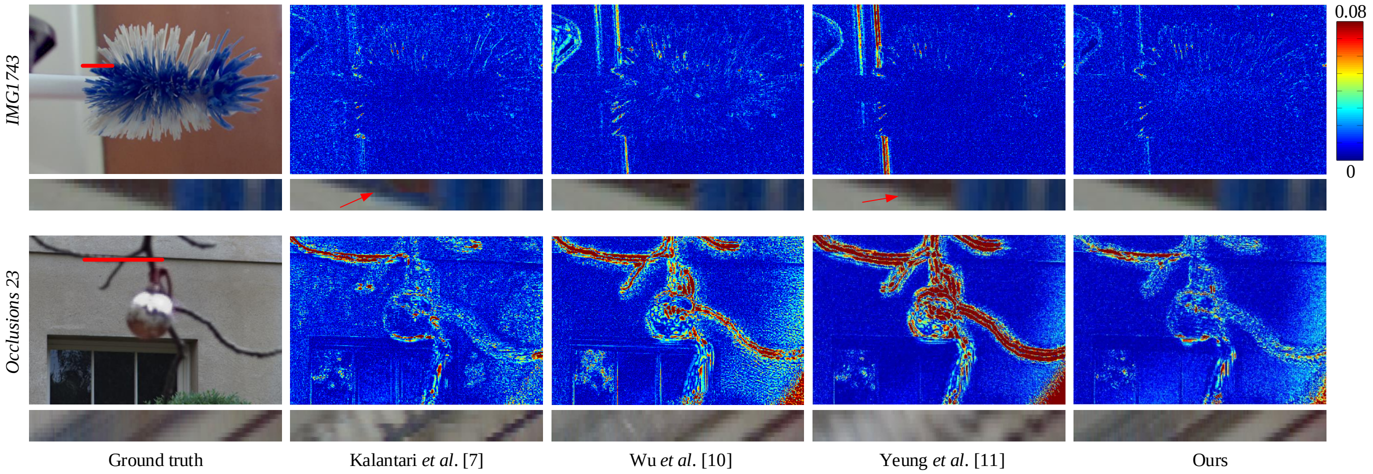


Fig. 9. Comparison of the results on the light fields from Lytro Illum. The results show the error map (absolute error of the grey-scale image) and the EPIs at the location marked by red lines. Light fields are from the *30 Scenes* [7] and the *Occlusions* category [57].

TABLE V
QUANTITATIVE RESULTS (PSNR/SSIM) OF RECONSTRUCTED VIEWS ON THE LIGHT FIELDS FROM LYTRO ILLUM [6]. THE *30 Scenes* DATASET COURTESY OF KALANTARI *et al.* [7], AND THE *Reflective* AND *Occlusions* CATEGORIES ARE FROM THE STANFORD LYTRO LIGHT FIELD ARCHIVE [57].

	Scale	*30 Scenes*	*Reflective*	*Occlusions*
Kalantari *et al.* [7]	3×	39.62/0.978	37.78/0.971	34.02/0.955
Wu *et al.* [10]		41.85/0.992	41.76/0.986	38.52/0.970
Yeung *et al.* [11]		44.53/0.990	42.56/0.975	39.27/0.945
Wang *et al.* [35]		43.82/0.993	39.93/0.959	34.69/0.923
Meng *et al.* [51]		- / -	40.14/0.964	36.05/0.929
Our proposed		**44.69/0.996**	**43.99/0.991**	**40.33/0.969**
Kalantari *et al.* [7]	7×	38.21/0.974	35.84/0.942	31.81/0.895
Wu *et al.* [10]		36.74/0.969	36.55/0.964	33.11/0.939
Yeung *et al.* [11]		**39.22**/0.977	36.47/0.947	32.68/0.906
Meng *et al.* [51]		- / -	36.97/ -	33.24/ -
Our proposed		39.09/**0.983**	**37.47/0.977**	**33.77/0.952**

light fields with higher view consistency (as shown in the demonstrated EPIs). Table V lists the quantitative results on the evaluated Lytro light fields. The PSNR and SSIM values are averaged over the light fields in each dataset.

VII. FURTHER ANALYSIS

A. Spatial-Angular Attention Map

We visualize additional attention maps on scenes with large disparity and non-Lambertian effect as shown in Fig. 10. We demonstrate the spatial-angular attention on a scene with large disparities in Fig. 10(a). In this case, the disparity between neighbouring views are about 16 pixels. Due to the spatial downsampling in the SAA-Net, the disparity of the light field (feature map) fed to the SAAM is about 4 pixels, see the top left figure in Fig. 10(a). Part of the attention map $M'(x_0, s_0, x_1, s_1), s_0 = 1, s_1 = 1, 2, 3$ is visualized in the bottom of Fig. 10(a). We can clearly see the the response moves from $R(A, A)$ at the positon $M'(13, 1, 13, 1)$ to $R(A, A'')$ at the position $M'(13, 1, 5, 1)$ along the angular dimension.

Fig. 10(b) demonstrates the spatial-angular attention on a scene with non-Lambertian effect. In this case, the positional relation of the corresponding points B, B' and B'' does not follow their depth, as clearly shown in the top right figure of Fig. 10(b). We visualize part of the attention map $M'(x_0, s_0, x_1, s_1), s_0 = 3, s_1 = 1, 2, 3$ in the bottom of Fig. 10(b). The result shows that the proposed SAAM is able to catch the correspondences even for regions with non-Lambertian effects.

B. Tensor Decomposition for Spatial-Angular Attention

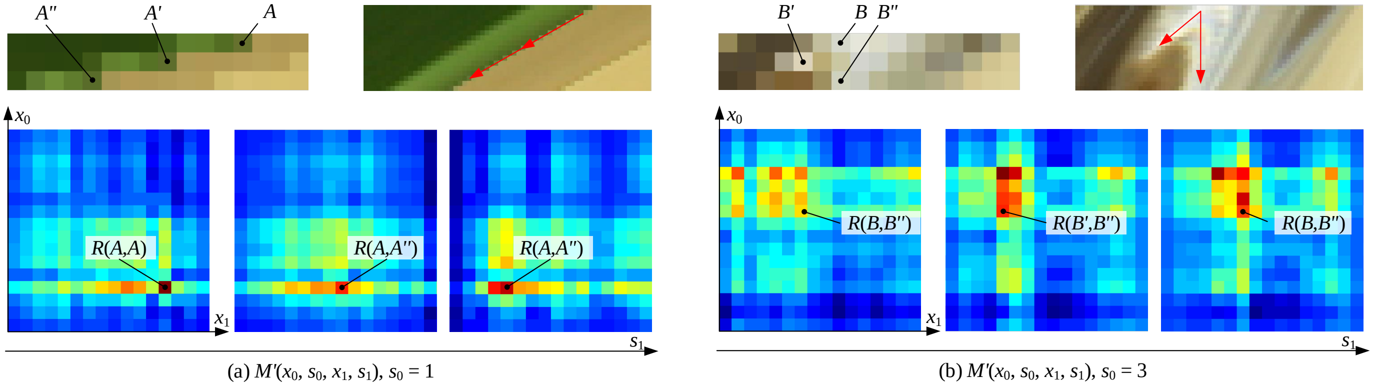
Although we propose a multi-scale reconstruction structure to alleviate the GPU memory cost, the SAAM will still eat up the GPU memory when dealing with an input light field with high spatial-angular resolution. For example, when reconstructing light fields from the MPI Light Field Archive [16] (spatial resolution 960 × 720), we have to disassemble the 3D data into sub-light fields of resolution 960 × 24 × 25 (width, height and angular). Our investigation shows that the disassembling will cause a negative effect on the reconstruction quality.

We therefore apply the truncated Singular Value Decomposition (SVD) [58] to compact the 3D tensor $\phi'_q \in \mathbb{R}^{BH \times WA \times C'}$ and $\phi'_k \in \mathbb{R}^{BH \times C' \times WA}$ before computing the attention map

$$\phi = USV^T,$$

where ϕ denotes ϕ'_q or ϕ'^T_k, U and V are two orthogonal matrices (ignoring the batch dimension), and S is a diagonal matrix with singular values along its diagonal. By truncating the diagonal matrix S with the largest τ singular values, we can get a good approximation $\phi \approx US_\tau V^T$ and also compress the 3D tensors. Since the rank of the matrices are $C' = 6$, the parameter of the truncated SVD $\tau = [1, 2, \cdots, C']$.

Fig. 11 shows the performance on PSNR in function of the SVD decomposition using the largest $\tau = [1, 2, \cdots, 6]$ singular values. As we can see from the curve "SVD", with no less than three singular values, the SVD decomposition will maintain the network performance without using fine-tuning. Moreover, since the decomposition enables us to feed the network with higher spatial resolution input, e.g., from



(a) $M'(x_0, s_0, x_1, s_1)$, $s_0 = 1$

(b) $M'(x_0, s_0, x_1, s_1)$, $s_0 = 3$

Fig. 10. Additional results of attention map on scenes with (a) large disparity and (b) non-Lambertian effect.

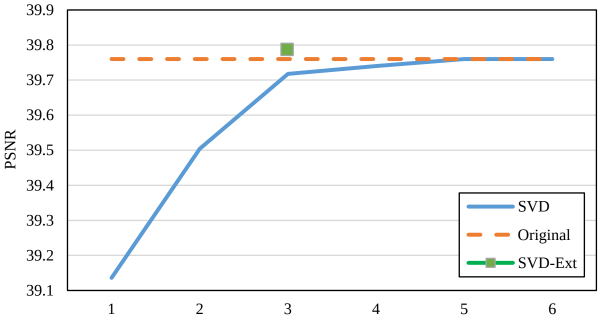


Fig. 11. The performance curve (PSNR) against the SVD decomposition of the proposed SAAM. The "SVD" denotes the truncated SVD with different parameters τ. The "Original" denotes the SAAM without SVD decomposition. The "SVD-Ext" denotes the truncated SVD with parameter $\tau = 3$ and the input sub-light fields of resolution $960 \times 64 \times 25$. The results are averaged on the 5 light fields from the MPI Light Field Archive [16].

$960 \times 24 \times 25$ to $960 \times 64 \times 25$, we can obtain a reconstruction result with even higher quality when employing truncated SVD decomposition, as shown by the "SVD-Ext" in the figure.

C. Limitations

The non-local attention involves outer product of large scale matrices, especially for the high-dimensional light field data. For this reason, the proposed network takes almost 15% of the time on the SAAM. For a 3D light field, the network takes about 53 seconds to reconstruct a 1×97 light field from 1×7 views of spatial resolution 960×720 ($16\times$ upsampling), i.e., 0.54s per view. For a 4D light field from Lytro Illum, it takes about 18 seconds to reconstruct a 7×7 light field from 3×3 views of spatial resolution 536×376 ($3\times$ upsampling), i.e., less than 0.36s per view. And the reconstruction of a 8×8 Lytro light field from 2×2 views ($7\times$ upsampling) takes less than 30 seconds, i.e., less then 0.5s per view. The above evaluations are performed on an Intel Xeon Gold 6130 CPU @ 2.10GHz with a NVIDIA Quadro GV100.

Although we apply a simple SVD decomposition to accerlate the network and compact the 3D tensor, the compression rate is limited by the rank of the matrices. Decomposing the attention map into the combination of small tensors [59] might solve this problem in a more essential way.

Repetitive patterns in the input light field can cause multiple plausible responses in the non-local attention, leading to misalignments in the reconstructed light fields. A possible solution is to introduce a smooth term in the attention map to penalize the multiple responses.

VIII. Conclusions

We have proposed a spatial-angular attention module in a 3D U-net backbone to capture correspondence information non-locally in the light field reconstruction problem. The introduced Spatial-Angular Attention Module (termed as SAAM) is designed to compute the responses from all the positions in the epipolar plane for each pixel in the light field and produce a spatial-angular attention map that records the correspondences. The attention map is then applied to driven the light field reconstruction via deconvolution in the angular dimension. We further propose a multi-scale reconstruction structure based on the 3D U-net backbone that implements the SAAM efficiently in the low spatial scale, while also preserving fine details in the high spatial scales by using decovlution-based reconstruction in each skip connection. For the network training, a spatial-angular perceptual loss is designed specifically for the high-dimensional light field data by pretraining a 3D auto-encoder. The evaluations on light fields with challenging non-Lambertian effects and large disparities have demonstrated the superiority of the proposed spatial-angular attention network.

References

[1] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
[2] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *ICML*, 2019, pp. 7354–7363.
[3] M. Levoy and P. Hanrahan, "Light field rendering," in *SIGGRAPH*, 1996, pp. 31–42.
[4] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, and P. Debevec, "A system for acquiring, processing, and rendering panoramic light field stills for virtual reality," in *SIGGRAPH*, 2019.
[5] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *European Conference on Computer Vision (ECCV)*, 2020.
[6] "Lytro." https://www.lytro.com/.
[7] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM TOG*, vol. 35, no. 6, 2016.
[8] J. Shi, X. Jiang, and C. Guillemot, "Learning fused pixel and feature-based view reconstructions for light fields," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *CVPRW*, 2015, pp. 24–32.

[10] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on EPI and extended applications," *IEEE TPAMI*, to be published.

[11] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," in *ECCV*, 2018, pp. 138–154.

[12] J. Long, Z. Ning, and T. Darrell, "Do convnets learn correspondence?" *Advances in Neural Information Processing Systems*, vol. 2, pp. 1601–1609, 2014.

[13] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *CVPR*, 2015.

[14] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[15] S. Moreschini, F. Gama, R. Bregovic, and A. Gotchev, "Civit dataset: Horizontal-parallax-only densely-sampled light-fields," https://civit.fi/densely-sampled-light-field-datasets/.

[16] V. K. Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a quality metric for dense light fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 58–67.

[17] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *NIPS*, 2016, pp. 658–666.

[18] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International conference on learning representations*, 2015.

[20] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *ICCV*, 2013, pp. 673–680.

[21] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *ICCV*, 2015, pp. 3487–3495.

[22] C.-T. Huang, "Robust pseudo random fields for light-field stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 11–19.

[23] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Computer Vision and Image Understanding*, vol. 145, pp. 148–159, 2016.

[24] E. Penner and L. Zhang, "Soft 3D reconstruction for view synthesis," *ACM TOG*, vol. 36, no. 6, p. 235, 2017.

[25] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[26] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE TPAMI*, vol. 36, no. 3, pp. 606–619, 2014.

[27] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75.

[28] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, "Attention-based view selection networks for light-field disparity estimation," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2020, p. 1.

[29] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *ECCV*, 2018.

[30] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," in *SIGGRAPH*, 2018.

[31] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 307–318.

[32] C. Zhang and T. Chen, "Spectral analysis for sampling image-based rendering data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 11, pp. 1038–1050, 2003.

[33] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE TPAMI*, vol. 40, no. 1, pp. 133–147, 2018.

[34] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous fourier domain," *ACM TOG*, vol. 34, no. 1, p. 12, 2014.

[35] Y. Wang, F. Liu, K. Zhang, Z. Wang, Z. Sun, and T. Tan, "High-fidelity view synthesis for light field imaging with extended pseudo 4DCNN," *IEEE Transactions on Computational Imaging*, pp. 1–1, 2020.

[36] H. Zhu, M. Guo, H. Li, Q. Wang, and A. Robleskelly, "Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–15, 2019.

[37] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional lstm network: a machine learning approach for precipitation nowcasting," in *NIPS*, 2015, pp. 802–810.

[38] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared epi structure for light field reconstruction," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3261–3273, 2019.

[39] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[40] R. A. Rensink, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[41] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[43] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys*, vol. 52, no. 1, p. 5, 2019.

[44] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[45] S. Woo, J. Park, J. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[46] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 250–12 259.

[47] K. H. Jin, M. T. Mccann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[48] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," in *International Conference on Learning Representations*, 2015.

[49] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *Acm Transactions on Graphics*, 2017.

[50] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7982–7991.

[51] N. Meng, H. K. So, X. Sun, and E. Y. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[52] "Stanford (New) Light Field Archive." http://lightfield.stanford.edu/lfs.html.

[53] Ng and Ren, "Fourier slice photography," *Acm Transactions on Graphics*, vol. 24, no. 3, p. 735, 2005.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

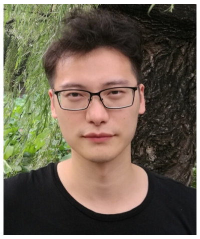
[55] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous systems." http://tensorflow.org/.

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

[57] "Stanford Lytro Light Field Archive." http://lightfields.stanford.edu/.

[58] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[59] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *Siam Review*, vol. 51, no. 3, pp. 455–500, 2009.



Gaochang Wu received the B.S. and M.S. degrees in mechanical engineering from Northeastern University, Shenyang, China, in 2013, and 2015, respectively. He is currently working toward Ph.D. degree in control theory and control engineering in Northeastern University, Shenyang, China. His current research interests include light field processing, computational photography and deep learning.

Yebin Liu received the BE degree from Beijing University of Posts and Telecommunications, China, in 2002, and the PhD degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He has been working as a research fellow at the computer graphics group of the Max Planck Institute for Informatik, Germany, in 2010. He is currently an associate professor in Tsinghua University. His research areas include computer vision and computer graphics.

Lu FANG is currently an Associate Professor at Tsinghua University. She received her Ph.D in Electronic and Computer Engineering from HKUST in 2011, and B.E. from USTC in 2007, respectively. Dr. Fang's research interests include image / video processing, vision for intelligent robot, and computational photography. Dr. Fang serves as TC member in Multimedia Signal Processing Technical Committee (MMSP-TC) in IEEE Signal Processing Society.

Tianyou Chai received the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 1985. He has been with the Research Center of Automation, Northeastern University, Shenyang, China, since 1985, where he became a Professor in 1988 and a Chair Professor in 2004. His current research interests include adaptive control, intelligent decoupling control, integrated plant control and systems, and the development of control technologies with applications to various industrial processes. Prof. Chai is a member of the Chinese Academy of Engineering, an academician of International Eurasian Academy of Sciences, IEEE Fellow and IFAC Fellow. He is a distinguished visiting fellow of The Royal Academy of Engineering (UK) and an Invitation Fellow of Japan Society for the Promotion of Science (JSPS).