Spatial-Angular Attention Network for Light Field Reconstruction

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

Abstract—Typical learning-based light field reconstruction methods demand in constructing a large receptive field by deepening their networks to capture correspondences between input views. In this paper, we propose a spatial-angular attention network to perceive non-local correspondences in the light field, 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 response of each query pixel from all the positions on the epipolar plane, and generate an attention map that captures correspondences along the angular dimension. Then a multi-scale reconstruction structure is proposed to efficiently implement the non-local attention in the low resolution feature space, while also preserving the high frequency components in the high-resolution feature space. 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

T THROUGH capturing both intensities and directions from sampled light rays, light field enables high-quality view synthesis without the need of complex and heterogeneous information such as geometry and texture. More importantly, benefiting from the light field rendering technology [3], photorealistic views can be rendered in real-time regardless of the scene complexity or non-Lambertian effect. This high quality rendering technology usually requires a densely-sampled light field (DSLF), where the disparity between adjacent views should be less than one pixel. 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. That is, due to the limitation of the sensor resolution [4], the light field is sparsely sampled either in the angular domain [5] or the spatial domain [6].

Recently, a more promising way is the fast capturing of a sparsely-sampled (angular domain) light field followed by direct reconstruction or depth-based view synthesis methods [7], [8] by using 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], these networks fail to collect enough information (i.e., the spatial-angular correspondences) when dealing with large disparities, leading to aliasing effects in the reconstructed light field. On the other hand, depth-based view synthesis methods [7], [13], [14] address the large disparity problem through plane sweep (depth estimation), and then synthesize novel views using learning-based prediction. However, these methods require depth consistency along the angular dimension, and thus, often fail to handle the depth ambiguity caused by the non-Lambertian effect.

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 corre-
spondences on the Epipolar Plane Image (EPI) in a non-local fashion, addressing the aforementioned non-Lambertian issue and large disparity issue in a unified framework. Specifically, the SAA-Net consists of two parts, including a spatial-angular attention module (Sec. IV-A) and a U-net backbone (Sec. IV-B). Motivated by the non-local attention mechanism in [11], for each pixel in the input light field, the Spatial-Angular Attention Module (termed as SAAM for short) computes the responses of pixels from all the positions on the epipolar plane, and generates 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 matrix multiplication and channel-to-angular pixel shuffling.

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 resolution feature space while preserving fine details in the high spatial resolution feature space.

For 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 into a 2D CNN (e.g., the commonly-used VGG [19]), we pre-train a 3D auto-encoder that considers the consistency in the network, we propose a multi-scale reconstruction structure 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 resolution feature space while preserving fine details in the high spatial resolution feature space.

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 challenging cases with both non-Lambertian effects and large disparities, as illustrated in Fig. 1 (bottom).

II. RELATED WORK

A. Light Field Reconstruction

First, we briefly review the major works on light field synthesis (or view synthesis) depending on whether the depth information is explicitly used.

1 The source code is available at https://github.com/GaochangWu/SAAN

Depth image-based view synthesis. Typically, these kind of approaches first estimate the depth of a scene, then warp and blend the input views to synthesize a novel view [20], [9]. Conventional light field depth estimation approaches follow the pipeline of stereo matching [21], i.e., cost computation, cost aggregation (or cost volume filtering), disparity regression and post refinement. The main difference is that light field converts disparity from discrete space into a continuous space [22], delivering various depth cues that enable light field depth estimation with different strategies, such as structure tensor-based local direction estimation [22], depth from correspondence [23], depth from defocus [24], [25] and depth from parallelogram cues [26]. Moreover, some learning-based approaches incorporate the aforementioned depth estimation pipeline with 2D convolution-based feature extraction, 3D convolution-based cost volume aggregation and depth regression [27]. 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 [22], soft blending [20] and learning-based synthesis [28].

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 using the predicted probabilities and colors for each depth plane. Kalantari et al. [17] further employed a sequential network setting to infer depth (disparity) and color, and optimized the model via end-to-end training. Following the sequential network setting, Meng et al. [29] developed a confidence estimation network between depth and color networks to infer pixel-wise blending weights. Shi et al. [8] proposed to blend the warped views in both pixel level and feature level. Jin et al. proposed to use a regular sampling pattern (four corner views) [30] and a flexible sampling pattern [31] for light field depth estimation, and then perform the reconstruction using spatial-angular alternating refinement. Ko et al. [32] introduced a dynamic blending filter to generate the filter coefficients adaptively according to the warped views. Different from the sequential network settings mentioned above, Zhou et al. [33] proposed a novel learning-based Multi-Plane Image (MPI) representation that infers a novel view by alpha blending of different images. Mildenhall et al. [14] further proposed to use multiple MPIs to synthesize a local light field.

Depth image-based view synthesis approaches solve the problem of large correspondence gap in the sparsely-sampled light field by using depth estimation and warping. But the scene depth are based on the Lambertian assumption, and thus, these approaches will suffer from depth ambiguity when addressing the non-Lambertian effect, as demonstrated in Fig. 8 (second case). In this paper, we address the problem of large correspondence gap with a non-local attention mechanism to capture large gap correspondences. Since we do not rely on depth information, the proposed method shows higher reconstruction quality on the non-Lambertian cases.

Reconstruction without explicit depth. These kind of approaches treat 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 [10]. Classical approaches [34], [35] 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. [36] utilized an adapted discrete shearlet transform in the Fourier domain to remove the high-frequency spectra that introduce aliasing effects. Shi et al. [37] performed DSLF reconstruction as an optimization for sparsity in the continuous Fourier domain.

Inspired by the success of deep learning in computer vision, depth-independent plenoptic reconstruction approaches [9], [10] were widely investigated by using deep convolution networks. Specifically, Zhu et al. [38] proposed an auto-encoder that combines convolutional layers and convLSTM layers [39]. For explicitly addressing the aliasing effects, Wu et al. [40] 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, this approach fails to recover the high-frequency details, and thus leading to blur effect. Liu et al. [40] further applied a multi-stream network that takes 3D EPIs in different directions as input. In addition to extracting slices from the plenoptic function, Yeung et al. [41] directly fed the entire 4D light field into a pseudo 4D convolutional network, and proposed a novel spatial-angular alternating convolution to iteratively refine the angular dimensions of the light field. Jin et al. [41] further extended the spatial-angular alternating convolution to the problem of compressive light field reconstruction. Wang et al. [42] applied pseudo 4D convolution to reconstruct the two angular dimensions of the input light field sequentially. Wu et al. [43] proposed an evaluation network for EPIs with different shear amount, termed as sheared EPI structure, and further boosted its performance with an end-to-end optimization framework [44]. With this structure, the networks implicitly use depth information to select a well reconstructed EPI. However, the performances of these networks are limited by the finite perceptive field of the convolutional neurons, especially when handling the large disparity problem.

B. Attention Mechanism

Attention was first built to imitate the mechanism of human perception that mainly focuses on the salient part [45], [46], [47]. Vaswani et al. [48] indicated that the attention mechanism is able to solve the long term dependency problem even without using convolution operation or recurrent neural cell. Attention mechanism is typically embedded within a conventional network backbone, such as a VGG-net [19], a ResNet [49], [50] or even a simple multi-layer perceptron (MLP) [51]. It encourages the network to focus on the salient parts by assigning an adaptive weight (attention map) to the extracted features. Hu et al. [52] and Woo et al. [50] proposed to use a global pooling (max-pooling or average-pooling) followed by an MLP to aggregate the entire information in the spatial dimension. This attention mechanism enables the network to focus on certain channels in the feature maps. However, the global pooling operation will decimate the high-frequencies in the feature maps [53], which could be unacceptable, especially for a reconstruction task. Alternatively, Vaswani et al. [48] proposed to use a weighted average of the responses from all the positions with respect to a certain position for Natural Language Processing (NLP), which is called self-attention. Wang et al. [1] further bridged the self-attention for NLP to more general tasks in computer vision, such as video classification. More concretely, a high-dimensional feature map, e.g., a 3D tensor with spatial-temporal dimensions, is reshaped into its 2D form for the operation of matrix multiplication. Different from the convolution, the self-attention allows the element in the feature map to interact with any elements regardless of their distance (range), and thus, is also termed as non-local attention.

Rather than using the non-local attention mechanism, Wang et al. [54], [55] proposed a parallax attention module to calculate the correspondence between two stereo images along the epipolar line. Tsai et al. [56] introduced an attention module in the angular dimension to weight the contribution of each view in a light field. Guo et al. [57] applied a pixel-wise attention map using convolution layers for adaptively blending feature maps from different frequency components in a light field.

Compared with the attention modules in [54], [55], [56], [57], the major difference in this paper is that the proposed attention is calculated non-locally in the 2D epipolar plane for each pixel, enabling the network to capture spatial-angular correspondences. That also makes us different from existing non-local attention mechanisms [1], [2], [48], which perform non-local attention across the entire data dimensions. Another significant difference is that we embed a channel-to-angular pixel shuffling operation into the non-local attention mechanism to explicitly achieve the light field reconstruction task. To the best of our knowledge, our method is the first to apply non-local attention to the light field reconstruction task.

III. Problem Analysis and Motivation

In this section, we empirically show that the performance of a learning-based light field reconstruction method is closely related to the perception range of its neurons (or convolutional filters) on the epipolar plane, especially when handling a light field with large disparity problem.

Deep neural network is proved to be a powerful technique in solving ill-posed inverse problems [58]. For the light field reconstruction task, both network structure and the disparity range of the scene are significant to the performance of light field reconstruction. Since the disparity range of the scene is unalterable once the light field is acquired, most deep learning-based light field reconstruction methods pursue a more appropriate architecture for better performance. Specifically, the depth-based view synthesis methods convert the feature maps into a physically meaningful depth map, while the depth-independent methods directly convert feature maps to novel views. Essentially, both two kinds of approaches adopt convolution operation to generate responses (feature maps) among corresponding pixels.
To quantitatively measure the capability of correspondence capturing, we apply the concept of receptive field introduced in \[59\]. 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 a certain convolutional filter.

We analyse the reconstruction qualities of two networks with the same structure (U-net) and the same number of parameters (around 120K) but different receptive fields, as illustrated in Fig. 2. For a scene with small disparity (about 3 pixels in the demonstrated example), networks with either small receptive field (27 × 27 pixels) or 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 corresponding pixels on the epipolar plane, as shown clearly at the top of Fig. 2(c). Since the actual size of the receptive field can be smaller than its theoretical size [59], the actual receptive field might not be able to cover the disparity range of the input light field, leading to severe aliasing effects in the reconstructed result, as shown at the bottom of Fig. 2(c). In contrast, the network with a large receptive field can produce high quality result (Fig. 2(d)).

Due to the limitation of parameter amount, it is intractable to expand the receptive field by pursuing a deeper network or a larger filter size. The fundamental idea of our proposed light field reconstruction method is to capture the correspondence non-locally across the spatial and angular dimensions of the light field. We achieve this with two features: 1) a spatial-angular attention module that captures non-local correspondence between any two pixels on the epipolar plane; and 2) an encoder-decoder network that reduces the redundancies in the light field to efficiently achieve the non-local perception.

IV. SPATIAL-ANGULAR ATTENTION NETWORK

In this section, we first introduce the overall architecture of the proposed SAA-Net for light field reconstruction, and then introduce the proposed spatial-angular attention module in details. 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 fields into 3D slices, the proposed network can be applied to both 3D light fields from a single-degree-of-freedom gantry system and 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 \(L_{x}(x, y, s)\) and \(L_{s}(y, x, t)\), then use the synthesized 3D light fields to reconstruct the final 4D light field.

A. Network Architecture

We propose a network with Multi-Scale Reconstruction (MSR) structure to maintain view consistency (i.e., continuity in the angular dimension) in the low spatial resolution feature space while preserving fine details in the high spatial resolution feature space. As shown in Fig. 3(a), the backbone of the proposed SAA-Net follows the encoder-decoder structure with skip connections (also known as U-net). But the proposed SAA-Net has two particular differences: 1) In each skip connection, we use a deconvolution layer along the angular dimension before feeding the feature maps to the decoder part; 2) In the encoder part, we use strided convolution with stride only 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 generates multi-scale light field features and reduces the redundant information in the spatial dimension to save the computational and GPU memory costs for the non-local perception. We use two convolutional layers (3D) with stride [2, 2] and [2, 1] to downsample the spatial resolution of the light field feature with ratio 2 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 \[11\].

The skip connections copy the feature layers before each downsampling layer in the encoder, as shown in Fig. 3(a). For each skip connection, a deconvolution layer (also known as transposed convolution layer) is applied to upsample the feature map in the angular dimension, followed by a \(1 \times 1 \times 1\) convolution. Since the angular information mainly concentrates on the 2D EPI \(E(x, s)\) for reconstructing a 3D light field \(L(x, y, s)\), the filter size in each deconvolution layer in the skip connection is set to \(3 \times 1 \times 7\).
and 3 use two 3D convolutional layers with filter sizes in the channel dimension \([60]\), as shown in Fig. 3(a). We then by concatenating the features from corresponding levels along the spatial dimensions (width and height), respectively. The reconstruction with pixel shuffling \([61]\) that disentangles the rather than the entire 3D space; 2) We model the light field attention mechanism is performed in the 2D epipolar plane disparity information is encoded in the EPI, the non-local differences between the proposed SAAM and the previous

**B. Spatial-Angular Attention Module**

Inspired by the non-local attention mechanism in \([1, 2]\), we propose a Spatial-Angular Attention Module (SAAM) to disentangle 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) We model the light field reconstruction with pixel shuffling \([61]\) that disentangles the reconstructed angular information from the channel dimension.

A straightforward choice to perform 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 due to the high computation complexity and GPU memory cost. Alternatively, we insert the proposed SAAM between the encoder and decoder, i.e., to perform non-local operation in the low-resolution feature space, as shown in Fig. 3.

Since features in a 3D CNN will be a 5D tensor \(\phi \in \mathbb{R}^{B \times W \times H \times A \times C'}\) (i.e., batch, width, height, angular and channel), we first apply two convolution layers with kernel size \(1 \times 1 \times 1\) to produce two feature layers \(\phi_q\) and \(\phi_k\) with size of \(B \times W \times H \times A \times C'\). The channel number \(C'\) is set to be \(\frac{C}{2}\) (i.e., \(C' = 6\) in our implementation) for computation efficiency. Then the feature layers \(\phi_q\) and \(\phi_k\) are reshaped into 3D tensors \(\phi'_q\) and \(\phi'_k\) of size \(BH \times WA \times C'\) 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\ on\ the\ epipolar\ plane.

We apply batch-wise matrix multiplication between \(\phi'_q\) and \(\phi'_k\) and use a softmax function to produce an attention map \(M\) as illustrated in Fig. 3(b). The attention map is composed of \(BHf\) matrices with shape \(WA \times WA\). Each matrix can be considered as a 2D expansion map of a 4D tensor \(M' \in \mathbb{R}^{W \times A \times H \times A}\) (the batch and height dimensions are neglected). The point \(M'(x_0, s_0, x_1, s_1)\) indicates the response of light
(a) An EPI with a foreground point $P_A$ and a background point $P_B$; (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.

**TABLE I**

| Layer   | $k$  | $s$ | $chn$ | Input        |
|---------|------|-----|-------|--------------|
| Encoder |      |     |       |              |
| Conv1_1 | $3 \times 1 \times 3$ | [1,1,1] | 1/24 | $L(x, y, s)$ |
| Conv1_2 | $1 \times 3 \times 3$ | 1,1,1 | 24/24 | Conv1_1     |
| Conv1_3 | $3 \times 3 \times 3$ | 2,2,1 | 24/48 | Conv1_2     |
| Conv2_1 | $3 \times 1 \times 3$ | 1,1,1 | 48/48 | Conv2_1     |
| Conv2_2 | $1 \times 3 \times 3$ | 1,1,1 | 48/48 | Conv2_2     |
| Conv2_3 | $3 \times 1 \times 3$ | 1,1,1 | 48/48 | Conv2_3     |
| Conv3_1 | $1 \times 3 \times 3$ | 1,1,1 | 96/48 | Conv3_1     |
| Conv3_2 | $3 \times 1 \times 3$ | 1,1,1 | 96/48 | Conv3_2     |
| Conv3_3 | $3 \times 1 \times 3$ | 1,1,1 | 96/48 | Conv3_3     |
| Conv3_4 | $1 \times 3 \times 3$ | 1,1,1 | 96/96 | Conv3_4     |
| Conv4_1 | $3 \times 1 \times 3$ | 1,1,1 | 96/96 | Conv4_1     |
| Conv5_1 | $3 \times 1 \times 3$ | 1,1,1 | 96/96 | Conv5_1     |
| Conv6_1 | $3 \times 1 \times 3$ | [1,1,1] | 1/1 | SAAM        |
| Conv7_1 | $4 \times 1 \times 1$ | [1,1,1] | 1/1 | Conv6_1     |
| SAAM    |      |     |       |              |
| Conv6_2 | $3 \times 1 \times 3$ | 1,1,1 | 48/48 | Conv6_2     |
| Conv6_3 | $3 \times 1 \times 3$ | 1,1,1 | 48/48 | Conv6_3     |
| Conv6_4 | $3 \times 1 \times 3$ | 1,1,1 | 48/48 | Conv6_4     |
| Conv7_2 | $4 \times 1 \times 1$ | 1,1,1 | 24/24 | Conv7_2     |
| Conv7_3 | $3 \times 1 \times 3$ | 1,1,1 | 24/24 | Conv7_3     |
| Conv8  | $3 \times 1 \times 3$ | 1,1,1 | 24/1  | Conv8       |

**Fig. 4.** Visualization of the attention map before the softmax function. (a) An EPI with a foreground point $P_A$ and a background point $P_B$; (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.

**Fig. 5.** We employ channel-to-angular pixel shuffling for light field reconstruction. To avoid view extrapolation effect in the reconstructed light field, we remove the last $\alpha - 1$ elements in the angular dimension.

$P_A$ will be occluded by the foreground point $P_B$, which is demonstrated more obviously in Fig. 4(b). Fig. 4(c) shows three sub-maps extracted from the attention map $M'$ with $s_0 = 1$ and $s_1 = 1, 2, 3$, respectively. It can be clearly seen that a point will have the highest response at the location of its correspondence in the attention map. For instance, the response $R(P_B, P_B')$ at the location $M'(11, 1, 9, 2)$ for the corresponding patch $(P_B, P_B')$ (the middle sub-figure of Fig. 4(c)), and the response $R(P_B, P_B')$ at the location $M'(11, 1, 7, 3)$ for the corresponding patch $(P_B, P_B')$ (the right sub-figure of Fig. 4(c)). For the occluded point $P_A$, the location of the maximum response changes from $M'(5, 1, 5, 1)$ to $M'(5, 1, 7, 3)$. In this case, the attention module is able to locate the occluded point $P_A'$ through its surrounding pixels. More demonstrations of spatial-angular attention map can be found in Sec. VII-A.

Feature $\phi_A$ and $\phi_B$ are obtained by another two $1 \times 1 \times 1$ convolutions in a similar manner as that for $\phi_A$ and $\phi_B$. The main difference is that the channel numbers of these two feature layers are $\alpha C'$, where, $\alpha$ denotes the reconstruction factor (upsampling scale in the angular dimension) of the network. Another batch-wise matrix multiplication is applied between the attention map $M'$ and $\phi_A'$ (reshaped from $\phi_A$), resulting a 3D tensor $\phi_A'' \in \mathbb{R}^{BH \times W \times A \times \alpha C'}$. We then reshape $\phi_A''$ into a 5D tensor $\phi_A''^\alpha / (H \times W \times A) \times \alpha C'$. Using the aforementioned SAAM, the reconstructed angular information can be well encoded in the channel dimension of the 5D tensor. By disentangling it with channel-to-angular pixel shuffling, we can reconstruct a high angular resolution light field (feature map) in a non-local manner. Specifically,
we first multiply the feature layer $\phi_a$ by a trainable scale parameter (initialized as 0) and add back to the feature layer $\phi_b$. We then apply the channel-to-angular pixel shuffling and reconstruct a 5D tensor $\phi_c \in \mathbb{R}^{B \times W \times H \times (A(A-1)+1) \times C}$. As illustrated in Fig. 3, the channel-to-angular pixel shuffling rearranges elements from channel dimension to angular dimension. The final output of the SAAM is generated by two convolutional layers with kernel sizes of $7 \times 1 \times 1$ and $1 \times 1 \times 7$, respectively.

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 the SAAM is listed in Table I.

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) [8]. Recently, researches [33], [62], [63], [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], [63], [29] 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 images, neglecting the view consistency constraint in the angular dimension. To achieve this, we design another 3D encoder-decoder network (auto-encoder) that learns to extract the high-level features for the spatial-angular perceptual loss. Note that the auto-encoder can be also generalized to a 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 its input, we only adopts 3D convolution in the encoder and decoder.

The proposed auto-encoder for the spatial-angular perceptual loss uses the output of the SAA-Net or the desired 3D light field (ground truth) as input, as shown in Fig. 6. It has 12 convolutional layers $(\phi_{ae}^{(l)}, l = 1, 2, \ldots, 12)$ with kernel size $3 \times 3 \times 3$ and 3 bilinear upsampling layers. The encoder part includes the first 6 convolutional layers, where 3 convolutional layers $(\phi_{ae}^{(2)}, \phi_{ae}^{(4)}, \phi_{ae}^{(6)})$ with stride 2 in each dimension are used to compress the light field from low-level pixel space into high-level feature space. The decoder part includes the last 6 convolutional layers and 3 upsampling layers to restore the light field from the latent representations. Detailed configuration of each layer (stride and number of channels) is shown in Fig. 6.

The auto-encoder learns how to extract high-level features through unsupervised learning, i.e., the network learns to predict the input 3D light field. The objective for training the auto-encoder is $\mathcal{L}_{AE}(L_{HR}) = \|f_{AE}(L_{HR}) - L_{HR}\|_1$, where $f_{AE}$ denotes the auto-encoder. We use the same training dataset (Sec. V-B), learning rate and optimizer (Sec. V-C) as those for the SAA-Net.

To compute the final spatial-angular perceptual loss, we feed the output of the SAA-Net $\hat{L}_{HR}$ as well as the ground truth light field $L_{HR}$ to the auto-encoder, as shown in Fig. 6. And the spatial-angular perceptual loss is defined as follows

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

where $\phi_{ae}^{(l)}(\cdot)$ (l = 2, 4, 6) denotes the feature maps obtained from the $l$th layer in the encoder, and $\lambda_{feat} = 0.2, 0.2, 0.1$ is a set of hyperparameters for the proposed spatial-angular perceptual loss.

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

$$\mathcal{L}_{SAA} = \mathcal{L}_{pix} + \mathcal{L}_{feat}.$$

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

B. Training Data

We use light fields from the Stanford (New) Light Field Archive [64] as the training dataset, which contains 12 light fields with $17 \times 17$ views. Since the network input is 3D light field, the light field Lego Gantry Self Portrait is excluded from the training dataset since the moving camera may influence the reconstruction performance.
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 [38], we augment the extracted 3D light fields using shearing operation [65]

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

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 \( d \). \( L_d(x, y, t) \) can be obtained following a similar manner. 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. 7 enabling the network to address the large disparity problem.

To accelerate the training procedure, the extracted 3D light fields are cropped into sub-light fields with a spatial resolution of \( 64 \times 24 \) (width and height for \( L(x, y, s) \)) or height and width for \( L(y, x, t) \) and a stride of 40 pixels. About \( 6.7 \times 10^5 \) examples can be extracted from the 3D light fields (original and sheared).

C. Implementation Details

We train two models with reconstruction factors \( \alpha = 3, 4 \). The input/output angular resolution of the training samples for these two models are 6/16 and 5/17, 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 a standard deviation of \( 1 \times 10^{-3} \), and the biases by zero. The network is optimized by using ADAM solver [66] with learning rate of \( 1 \times 10^{-4} \) (\( \beta_1 = 0.9, \beta_2 = 0.999 \)) and mini-batch size of 28. The training model is implemented in the TensorFlow framework [67]. The network converges after \( 8 \times 10^5 \) steps of backpropagation, taking about 35 hours on a NVIDIA Quadro GV100.

VI. EVALUATIONS

In this section, we evaluate the proposed SAA-Net on various kinds of light fields, including those from both gantry systems and from plenoptic camera (Lytro Illum [4]). We mainly compare our method with six state-of-the-arts learning-based methods including Kalantari et al. [7] (depth-based), LLFF [14] (MPI representation), Wu et al. [10], Yeung et al. [11], HDDRNet [63] and DA^2N [44] (without explicit depth). To fully demonstrate the effectiveness of our design choices, we also perform ablation studies by training our network without the SAAM, without the MSR structure and without the spatial-angular perceptual loss, respectively. The quantitative evaluations is reported by measuring the average PSNR and SSIM [68] values over the synthesized views of the luminance channel in the YCbCr space. Please refer to the submitted video for more qualitative results.

A. Evaluations on Light Fields from Gantry Systems

In this experiment, the comparisons are performed on light fields from the MPI Light Field Archive [16] (1 \( \times \) 101 views of resolution 960 \( \times \) 720, 1 \( \times \) 97 views for evaluation) and the CIVIT Dataset [15] (1 \( \times \) 193 views of resolution 1280 \( \times \) 720) with upsampling scales 8\( \times \) and 16\( \times \). The performances with respect to both angular sparsity and non-Lambertian are taken into consideration. Since the vanilla version of the network by Yeung et al. [11] and Meng et al. [63] (HDDRNet) were specifically designed for 4D light fields, we modify their convolutional layers to fit the 3D input while keeping its network architecture unchanged. The networks by Kalantari et al. [7], Yeung et al. [11], Mildenhall et al. [14] (LLFF) and Meng et al. [63] (HDDRNet) are re-trained using the same training dataset as our SAA-Net. Due to the particularity of the training datasets in [44], we do not retrain the DA^2N. We perform network cascade to achieve different upsampling scales, i.e., two cascades for \( 8 \times \) (16\( \times \)) upsampling using a network of reconstruction factor \( \alpha = 3 \) (\( \alpha = 4 \)).

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\( \times \) (disparity range up to 33.5px). The first and the third cases have complex occlusion structures, as shown in the top and the bottom row in Fig. 8. The baseline methods [7], [10], [11], [14] fail to reconstruct the complex structures. Among them, the depth and learning-based approach [7] and the MPI-based approach [14] fail to estimate proper occlusion relations between the foregrounds and the backgrounds. The second scene is a non-Lambertian case, i.e., a refractive glass before the toys. The approach by Kalantari et al. [7] cannot reconstruct the refractive object. And the EPIs reconstructed by the baseline methods [10], [11] appear severe aliasing effects due to the limited receptive field. The MPI-based approach LLFF [14] is able to reconstruct the non-Lambertian effects in this case, but produces ghosting artifacts as marked by the red arrow in Fig. 8.

Fig. 9 shows the reconstruction results on three light fields, Castle, Holiday and Flowers, from the CIVIT Dataset [15] with upsampling scale 16\( \times \) (disparity range about 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] fails to estimate depth maps accurate enough to warp the input images, and the color CNN cannot correct the misaligned views, producing ghosting artifacts as shown in Fig. 9. For the second case, we demonstrate

\[d = 2\]

\[d = -2\]

Fig. 7. An illustration of training data augmentation using shearing operation. For clear display, one of the spatial dimension in the 3D light field is ignored.

3In 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 \( \alpha = 3 \) (\( \alpha = 4 \)).
reconstructed EPIs in a highly non-Lambertian region. 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. In the demonstrated cases, LLFF [14] assigns wrong planes to the tiny structures, leading to ghosting and tearing artifacts.

Table III and Table IV list the quantitative measurements (8× and 16× upsampling scales) on the light fields from the MPI Light Field Archive [16] and the CIVIT Dataset [15], respectively. Compared with the baseline approaches, the proposed SAA-Net shows superior performance on both light field datasets.

### 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 [69]. 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). We compare our SAA-Net with 8 learning-based approaches, including 3 depth-based approaches (Kalantari et al. [7], LLFF [14] and Meng et al. [29]) and 5 approaches without explicit depth (Wu et al. [10], Wang et al. [42], Yeung et al. [11], HDDRNet [63] and DA³N [44]). Since the vanilla versions of the networks in [7], [11], [44], [39], [29] are trained on Lytro light fields, we use their open-source model (parameters) without re-training. The networks by Wu et al. [10] and Mildenhall et al. [14] (LLFF) are re-trained using the same dataset introduced in Sec. V-B. 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 3× and 7× upsampling scales.

---

**TABLE III**

| Scale | Bikes | Fairy Collection | Living Room | Mannequin | Workshop | Average |
|-------|-------|------------------|-------------|-----------|----------|---------|
| **Kalantari et al. [7]** | 34.83 / 0.969 | 36.66 / 0.977 | 46.35 / 0.991 | 40.62 / 0.983 | 38.66 / 0.986 | 39.32 / 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 |
| LLFF [14] | 36.84 / 0.985 | 40.14 / 0.989 | 46.85 / 0.990 | 43.23 / 0.989 | 41.79 / 0.991 | 41.77 / 0.989 |
| HDDRNet [63] | 38.97 / 0.994 | 40.20 / 0.994 | 43.65 / 0.997 | 42.21 / 0.996 | 40.19 / 0.997 | 42.41 / 0.996 |
| DA³N [44] | 39.14 / 0.992 | 41.00 / 0.993 | 46.30 / 0.996 | 44.17 / 0.996 | 43.09 / 0.996 | 42.74 / 0.994 |
| **Our proposed** | 40.35 / 0.994 | 42.22 / 0.995 | 48.01 / 0.997 | 44.99 / 0.996 | 45.29 / 0.997 | 44.17 / 0.987 |
| Kalantari et al. [7] | 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 / 0.998 | 37.21 / 0.982 |
| LLFF [14] | 34.95 / 0.963 | 34.01 / 0.966 | 44.73 / 0.987 | 39.92 / 0.985 | 37.61 / 0.985 | 38.24 / 0.977 |
| HDDRNet [63] | 33.97 / 0.976 | 35.08 / 0.979 | 44.83 / 0.997 | 40.60 / 0.993 | 38.54 / 0.992 | 38.60 / 0.987 |
| DA³N [44] | 35.79 / 0.984 | 36.23 / 0.981 | 45.91 / 0.996 | 40.83 / 0.992 | 40.11 / 0.994 | 39.77 / 0.990 |
| **Our proposed** | 36.54 / 0.987 | 37.62 / 0.987 | 47.25 / 0.997 | 42.08 / 0.994 | 40.55 / 0.994 | 40.81 / 0.992 |
| MSR structure-2 | 34.15 / 0.972 | 35.27 / 0.974 | 46.21 / 0.996 | 40.97 / 0.993 | 38.66 / 0.991 | 39.14 / 0.987 |
| MSR structure-3 | 34.20 / 0.973 | 35.53 / 0.984 | 46.79 / 0.996 | 41.04 / 0.993 | 38.85 / 0.991 | 39.28 / 0.987 |
| MSR structure-4 | 34.45 / 0.975 | 35.05 / 0.977 | 47.00 / 0.995 | 41.01 / 0.993 | 38.26 / 0.991 | 39.15 / 0.987 |
| w/o MSR structure | 33.76 / 0.975 | 34.35 / 0.976 | 46.07 / 0.995 | 40.19 / 0.991 | 38.05 / 0.992 | 38.48 / 0.986 |
| w/o SAP loss | 36.43 / 0.987 | 37.29 / 0.986 | 47.10 / 0.997 | 41.81 / 0.993 | 40.18 / 0.993 | 40.56 / 0.991 |

"MSR structure-X" denotes the proposed network that has X series of the MSR structure without using the SAAM.
Table IV lists the quantitative results on the evaluated Lytro light fields. The proposed SAA-Net shows competitive performance compared with the state-of-the-art light field reconstruction approach by Yeung et al. [11]. For the evaluated 113 light fields, the average PSNR / SSIM values of the proposed network are 42.55 / 0.984 for 3× upsampling and 36.55 / 0.969 for 7× upsampling. In comparison, the average PSNR / SSIM values of the baseline approaches with the highest performance are 41.82 / 0.984 (DA²N [14]) for 3× upsampling and 35.81 / 0.938 (Meng et al. [29]) for 7× upsampling. Since our network is not re-trained or fine-tuned on any Lytro light field dataset, these experimental results clearly demonstrate that our network can generalize well on light fields captured by different acquisition geometries.

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 [69], as shown in Fig. 10. 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) due to their limited receptive fields. The depth and learning-based approaches by Kalantari et al. [7] and Mildenhall et al. [14] (LLFF) produce promising results in the first case, but appear tearing artifacts near the occlusion boundaries as marked by the red arrows in the EPIs. In the second case, the approach by Kalantari et al. [7] fails to estimate proper depth information, introducing misalignment as shown by the EPI. LLFF [14] produces ghosting effects due to the incorrect plane assignments around the background region, as shown by the red arrows in the figure. In comparison, the proposed SAA-Net provides reconstructed light fields with higher view consistency (as shown in the demonstrated EPIs).

### C. Ablation studies

In this experiment, we empirically validate the modules and losses in our SAA-Net by performing the following ablation studies on different datasets. The results are listed in the last
five rows in Table [V], [II] and [V]. First, we evaluate the network with increasing series of the MSR structure while removing the SAAM. We use “MSR structure-X” to represent the network with X series of MSR structure. By increasing the series, the network will have a larger receptive field, e.g., the theoretical receptive field size of the MSR structure-4 reaches 247 pixels in the spatial dimension. However, since it is intractable to increase the actual size of the receptive field simply by using a deeper network [59], purely increasing the series of MSR structure has limitations in improving performance. This point is also verified by the ablation study. In comparison, the non-local attention is more effective than simply increasing the receptive field of the network.

In the second ablation study, we use a typical 3D U-net with the same convolutional layers as the backbone and remove the deconvolution layer in the skip connections, denoted as “w/o MSR structure” for short. The angular reconstruction is simply achieved by using deconvolution at the end of the network. The network parameters are kept comparable to the SAA-Net by adjusting the channel dimension of the network. For light fields from gantry systems [16], [15], the performance of the network decreases more than 2dB in terms of PSNR, as shown in Table [V] and [II]. For light fields from Lytro Illum, the performance of the network decreases more than 1.3dB, as shown in Table [V].

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 around 0.3dB as shown in Table [V].

In Fig. [11], we also visualize the comparison (16x upsampling) between the proposed SAA-Net, the network with different series of the MSR structure while removing the SAAM and the network without the MSR structure. The evaluated light fields are from the MPI Light Field Archive [16]. The results show that without using the proposed modules, especially the MSR structure, the SAA-Net appears severe aliasing effects around region with large disparity, e.g., the occluded leaves in the FairyCollection case and the hand writing on the reflective board in the Mannequin case.

VII. FURTHER ANALYSIS

A. Spatial-Angular Attention Map

We visualize four additional attention maps (before the softmax function) on scenes with small disparity, large disparity, occlusion and non-Lambertian effect, as shown in Fig. [12]. The first case (Fig. [12]a) shows a scene with a relatively small disparity (about 7 pixels). Due to the spatial downsampling operation in the MSR structure, the disparity of the light field features in the SAAM is about 1.75 pixels, as shown in the top left figure in Fig. [12]a. Three submaps $M'(x_0, s_0, x_1, s_1), s_0 = 2, s_1 = 1, 2, 3$ are shown in the bottom of Fig. [12]a, which visualize the correspondence captured by the attention mechanism. As we can see, the response with the highest value moves from $R(P_A', P_A)$ at position $M'(11, 2, 12, 1)$ to $R(P_A', P_A)$ at position $M'(11, 2, 9, 3)$ along the angular dimension. In addition, the attention map also shows a high response value $R(P_A', P_B)$ at position $M'(11, 2, 13, 1)$ due to the fractional disparity. This indicates
that the proposed SAAM has potential to capture a correspondence with sub-pixel accuracy.

In the second case (Fig. 12(b)), we demonstrate the spatial-angular attention on a scene with a large disparity (about 16 pixels). The spatial downsampling operation in the MSR structure reduces the disparity to 4 pixels. The response in the attention map moves from \( R(P'_A, P_A) \) at position \( M'(10, 2, 14, 1) \) to \( R(P'_A, P'_A) \) at position \( M'(10, 2, 6, 3) \) along the angular dimension. Although the SAAM only applies \( 1 \times 1 \times 1 \) convolutions before the attention, it is capable to capture correspondence with large displacement.

Fig. 12(c) visualize the attention map of a scene with a complex occlusion structure (Mannequin in the MPI Light Field Archive [16]). The background white board is occluded by the foreground leaves (please refer to the sub-aperture image of the second row in Fig. 11). As shown by the demonstrated attention map in Fig. 12(c), the SAAM fails to capture a correct correspondence because of the occlusion. The highest response value is \( R(P'_A, P_B) \) at position \( M'(8, 2, 10, 1) \), which is also a background point like \( P_A \). However, the SAAM also generate a considerable response \( R(P'_A, P_A) \) at position \( M'(8, 2, 8, 1) \), which we speculate is inferred from other non-occluded background points.

The fourth case in Fig. 12(d) demonstrates the spatial-angular attention on a scene with non-Lambertian effect. In this case, the positional relation between the corresponding points \( P_A, P'_A \) and \( P'_A \) does not follow a clear depth cue, as clearly shown in the top right figure of Fig. 12(d). Analo-
can get a good approximation sub-light fields of resolution 960 × 720 (16× upsampling), i.e., 0.53s per view. For the 3D light field from the MPI Light Field Archive [16], the network takes about 126 seconds to reconstruct a 1 × 193 light field from 1 × 13 views of spatial resolution 1280 × 720 (16× upsampling), i.e., 0.65s per view. For a 4D light field from Lytro Illum, it takes about 17 seconds to reconstruct a 7 × 7 light field from 3 × 3 views of spatial resolution 536 × 376 (3× upsampling), i.e., about 0.35s per view. And the reconstruction of a 8 × 8 Lytro light field from 2 × 2 views (7× upsampling) takes about 30 seconds, i.e., about 0.5s per view. The parameter number of our SAA-Net is about 338K. The above evaluations are performed on an Intel Xeon Gold 6130 CPU @ 2.10GHz with an NVIDIA TITAN Xp.

Although we apply a simple SVD decomposition to accelerate 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 [71] might solve this problem in a more essential way.

The another limitation of our proposed method is that repetitive patterns in the input light field can cause multiple plausible responses in the attention map, leading to misalignments in the reconstructed light fields. A possible solution is to introduce a smooth term in the attention map as in [54] to penalize multiple responses.

VIII. CONCLUSIONS

We have proposed a spatial-angular attention module in a 3D U-net backbone to capture correspondence information non-locally for light field reconstruction. The introduced Spatial-Angular Attention Module (termed as SAAM) is designed to compute the responses from all the positions on 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 guide light field reconstruction via channel-to-angular pixel shuffling. We further propose a multi-scale reconstruction structure based on the 3D U-net backbone that implements the SAAM efficiently in the low spatial resolution feature space, while also preserving fine details in the high spatial resolution feature space. 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.
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. 12250–12259.

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.

M. Guo, J. Hou, J. Jin, J. Chen, and L.-P. Chau, “Deep spatial-angular regularization for compressive light field reconstruction over coded apertures,” in European Conference on Computer Vision, 2020, pp. 278–294.

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.

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.

G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, “Hdri image reconstruction from a single exposure using deep cnns,” Acm Transactions on Graphics, 2017.

W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.

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.

N. Meng, H. K.-H. 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, vol. 43, no. 3, pp. 873–886, 2021.

“Stanford (New) Light Field Archive.” http://lightfield.stanford.edu/lsf.html.

Ng and Ren, “Fourier slice photography,” Acm Transactions on Graphics, vol. 24, no. 3, p. 735, 2005.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Computer Science, 2014.

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/.

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.

“Stanford Lytro Light Field Archive.,” http://lytrofields.stanford.edu/.

http://lightfields.stanford.edu.

T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” Siam Review, vol. 51, no. 3, pp. 455–500, 2009.

Yingjian Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. degree in information and communication engineering from National University of Defense Technology (NUDT), Changsha, China, in 2018. He is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology, NUDT. His research interests focus on low-level vision, particularly on light field imaging and image super-resolution.

Lu FANG is currently an Associate Professor at Tsinghua University. She received her Ph.D degree 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 (MMS-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).