Light Field Image Super-Resolution Using Deformable Convolution

Yingqian Wang, Jungang Yang, *Member, IEEE*, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo, *Senior Member, IEEE*

***Abstract*— Light field (LF) cameras can record scenes from multiple perspectives, and thus introduce beneficial angular information for image super-resolution (SR). However, it is challenging to incorporate angular information due to disparities among LF images. In this paper, we propose a deformable convolution network (i.e., LF-DFnet) to handle the disparity problem for LF image SR. Specifically, we design an angular deformable alignment module (ADAM) for feature-level alignment. Based on ADAM, we further propose a collect-and-distribute approach to perform bidirectional alignment between the center-view feature and each side-view feature. Using our approach, angular information can be well incorporated and encoded into features of each view, which benefits the SR reconstruction of all LF images. Moreover, we develop a baseline-adjustable LF dataset to evaluate SR performance under different disparity variations. Experiments on both public and our self-developed datasets have demonstrated the superiority of our method. Our LF-DFnet can generate high-resolution images with more faithful details and achieve state-of-the-art reconstruction accuracy. Besides, our LF-DFnet is more robust to disparity variations, which has not been well addressed in literature.**

***Index Terms*— Light field, super-resolution, deformable convolution, dataset.**

I. Introduction

ALTHOUGH light field (LF) cameras enable many attractive functions such as post-capture image editing [1]–[3], depth sensing [4]–[9], saliency detection [10]–[14], and de-occlusion [15]–[17], the resolution of a sub-aperture image (SAI) is much lower than that of the total sensors. The low spatial resolution problem hinders the development of LF imaging [18]. Since high-resolution (HR) images are required in various LF applications, it is necessary to reconstruct HR images from low-resolution (LR) observations, namely, to perform LF image super-resolution (SR).

To achieve high SR performance, information both within a single view (i.e., spatial information) and among different views (i.e., angular information) is important. Several models have been proposed in early LF image SR methods, such as variational model [19], Gaussian mixture model [20], and PCA analysis model [21]. Although different delicately handcrafted image priors have been investigated in these traditional methods [19]–[24], their performance is relatively limited due to their inferiority in spatial information exploitation. In contrast, recent deep learning-based methods [25]–[31] enhance spatial information exploitation via cascaded convolutions, and thus achieve improved performance as compared to traditional methods. Yoon *et al.* [25], [26] proposed the first CNN-based method LFCNN for LF image SR. Specifically, sub-aperture images (SAIs) are first super-resolved using SRCNN [32], and then fine-tuned in pairs to incorporate angular information. Similarly, Yuan *et al.* [27] super-resolved each SAI separately using EDSR [33], and then proposed an EPI-enhancement network to refine the results. Although several recent deep learning-based methods [28]–[31] have been proposed to achieve the state-of-the-art performance, the *disparity* issue in LF image SR is still under-investigated.

In real-world scenes, objects at different depths have different disparity values in LF images. Existing CNN-based LF image SR methods [25]–[31] do not explicitly address the disparity issue. Instead, they use cascaded convolutions to achieve a large receptive field to cover the disparity range. As demonstrated in [34], [35], it is difficult for SR networks to learn the non-linear mapping between LR and HR images under complex motion patterns. Consequently, the misalignment impedes the incorporation of angular information and leads to performance degradation. Therefore, specific mechanisms should be designed to handle the disparity problem in LF image SR.

Inspired by the success of deformable convolution [36], [37] in video SR [38]–[42], in this paper, we propose a deformable convolution network (namely, LF-DFnet) to handle the disparity problem for LF image SR. Specifically, we design an angular deformable alignment module (ADAM) and a collect-and-distribute approach to achieve feature-level alignment and angular information incorporation. In ADAM, all side-view features are first aligned with the center-view feature to achieve feature collection. These collected features

Manuscript received April 11, 2020; revised September 2, 2020 and November 1, 2020; accepted November 25, 2020. Date of publication December 8, 2020; date of current version December 14, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61972435, Grant 61401474, and Grant 61921001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ajmal S. Mian. *(Corresponding authors: Jungang Yang; Yulan Guo.)*

Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, and Wei An are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: wangyingqian16@nudt.edu.cn; yangjungang@nudt.edu.cn; wanglongguang15@nudt.edu.cn; yingxinyi18@nudt.edu.cn; wutianhao16@nudt.edu.cn; anwei@nudt.edu.cn).

Yulan Guo is with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China, and also with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: yulan.guo@nudt.edu.cn).

Digital Object Identifier 10.1109/TIP.2020.3042059

are then fused and distributed to their corresponding views by performing alignment with their original features. Through feature collection and distribution, angular information can be incorporated and encoded into each view. Consequently, the SR performance is evenly improved among different views. Moreover, we develop a novel LF dataset named NUDT to evaluate the performance of LF image SR methods under different disparity variations. All scenes in our NUDT dataset are rendered using the 3dsMax software[1] and the baseline of virtual camera arrays is adjustable. In summary, the main contributions of this paper are as follows:

- We propose an LF-DFnet to achieve the state-of-the-art LF image SR performance by addressing the disparity problem.
- We propose an angular deformable alignment module and a collect-and-distribute approach to achieve high-quality reconstruction of each LF image. Compared to [28], our approach avoids repetitive feature extraction and can exploit angular information from all SAIs.
- We develop a novel NUDT dataset by rendering synthetic scenes with adjustable camera baselines. Experiments on the NUDT dataset have demonstrated the robustness of our method with respect to disparity variations.

The rest of this paper is organized as follows: In Secion II, we briefly review the related work. In Section III, we introduce the architecture of our LF-DFnet in details. In Section IV, we introduce our self-developed dataset. Experimental results are presented in Section V. Finally, we conclude this paper in Section VI.

II. Related Work

In this section, we briefly review the major works in single image SR (SISR), LF image SR, and deformable convolution.

A. Single Image SR

The task of SISR is to generate a clear HR image from its blurry LR counterpart. Since an input LR image can be associated to multiple HR outputs, SISR is a highly ill-posed problem. Recently, several surveys [43]–[45] have been published to comprehensively review SISR methods. Here, we only describe several mile-stone works in literature.

Since Dong *et al.* [32], [46] proposed the seminal work of CNN-based SISR method SRCNN, deep learning-based methods have dominated this area due to their remarkable performance in terms of both accuracy and efficiency. By far, various networks have been proposed to continuously improve the SISR performance. Kim *et al.* [47] proposed a very deep SR network (i.e., VDSR) and achieved a significant performance improvement over SRCNN. Lim *et al.* [33] proposed an enhanced deep SR network (i.e., EDSR). With the combination of local and global residual connections, EDSR won the NTIRE 2017 SISR challenge [48]. Zhang *et al.* [49], [50] proposed a residual dense network (i.e., RDN), which achieved a further improvement over the state-of-the-arts at that time. Subsequently, Zhang *et al.* [51] proposed a residual channel attention network (i.e., RCAN) by introducing a channel attention module and a residual in residual mechanism. More recently, Dai *et al.* [52] proposed SAN by applying the second-order attention mechanism to SISR. Note that, RCAN [51] and SAN [52] achieve the state-of-the-art SISR performance to date in terms of PSNR and SSIM.

In summary, SISR networks are becoming increasingly deep and complicated, resulting in continuously improved capability in spatial information exploitation. Note that, performing SISR on LF images is a straightforward scheme to achieve LF image SR. However, the angular information is discarded in this scheme, resulting in limited performance.

B. LF Image SR

In the area of LF image SR, both traditional and deep learning-based methods are widely used. For traditional methods, various models have been developed for problem formulation. Wanner and Goldluecke [19] proposed a variational method for LF image SR based on the estimated depth information. Mitra and Veeraraghavan [20] encoded LF structure via a Gaussian mixture model to achieve depth estimation, view synthesis, and LF image SR. Farrugia *et al.* [21] decomposed HR-LR patches into subspaces and proposed a linear subspace projection method for LF image SR. Alain *et al.* proposed LFBM5D for LF image denoising [53] and LF image SR [24] by extending BM3D filtering [54] to LFs. Rossi and Frossard [22] developed a graph-based method to achieve LF image SR via graph optimization. Although the LF structure is well encoded by these models [19]–[22], [24], the spatial information cannot be fully exploited due to the poor representation capability of these handcrafted image priors.

Recently, deep learning based SISR methods are demonstrated superior to traditional methods in spatial information exploitation. Inspired by these works, recent LF image SR methods adopted deep CNN to improve their performance. In the pioneering work LFCNN [25], [26], SAIs were first separately super-resolved via SRCNN, and then fine-tuned in pairs to enhance both spatial and angular resolution. Subsequently, Yuan *et al.* [27] proposed LF-DCNN to improve LFCNN by super-resolving each SAI via a more powerful SISR network EDSR and fine-tuning the initial results using a specially designed EPI-enhancement network. Apart from these two-stage SR methods, a number of one-stage network architectures have been designed for LF image SR. Wang *et al.* proposed a bidirectional recurrent network LFNet [29] by extending BRCN [55] to LFs. Zhang *et al.* [28] proposed a multi-stream residual network resLF by stacking SAIs along different angular directions as inputs to super-resolve the center-view SAI. Yeung *et al.* [30] proposed LFSSR to alternately shuffle LF features between SAI pattern and macro-pixel image pattern for convolution. More recently, Jin *et al.* [56] proposed an all-to-one LF image SR method (i.e., LF-ATO) and performed structural consistency regularization to preserve the parallax structure among reconstructed views. Wang *et al.* [31] proposed an LF-InterNet to interact spatial and angular information for LF image SR. LF-ATO and LF-InterNet are state-of-the-art LF image SR methods to date and can achieve a high reconstruction accuracy.

[1] https://www.autodesk.eu/products/3ds-max/overview

Although the performance is continuously improved by recent networks, the disparity problem has not been well addressed in literature. Several methods [25]–[28] use stacked SAIs as their inputs, making pixels of same objects vary in spatial locations. In LFSSR [30] and LF-InterNet [31], LF features are organized into a macro-pixel image pattern to incorporate angular information. However, pixels can fall into different macro-pixels due to the disparity problem. In summary, due to the lack of the disparity handling mechanism, the performance of these methods degrade when handling scenes with large disparities. Note that, LFNet [29] achieves LF image SR in a video SR framework and implicitly addresses the disparity issue via recurrent networks. Although all angular views can contribute to the final SR performance, the recurrent mechanism in LFNet [29] only takes SAIs from the same row or column as its inputs. Therefore, the angular information in LFs cannot be efficiently used.

C. Deformable Convolution

The fixed kernel configuration in regular CNNs hinders the exploitation of long-range information. To address this problem, Dai *et al.* [36] proposed deformable convolution by introducing additional offsets, which can be learned adaptively to make the convolution kernel process feature far away from its local neighborhood. Deformable convolutions have been applied to both high-level vision tasks [36], [57]–[59], and low-level vision tasks such as video SR [38]–[41]. Specifically, Tian *et al.* [38] proposed a temporal deformable alignment network (i.e., TDAN) by applying deformable convolution to align input video frames without explicit motion estimation. Wang *et al.* [39] proposed an enhanced deformable video restoration network (i.e., EDVR) by introducing a pyramid, cascading and deformable alignment module to handle large motions between frames. EDVR won the NTIRE19 video restoration and enhancement challenges [60]. More recently, deformable convolution is integrated with non-local operation [40], convolutional LSTM [41] and 3D convolutions [42] to further enhance the video SR performance.

In summary, existing deformable convolution-based video SR methods [38]–[42] only perform unidirectional alignments to align neighborhood frames to the reference frame. However, in LF image SR, it is computational expensive to repetitively perform unidirectional alignments for each view to super-resolve all LF images. Consequently, we propose a collect-and-distribute approach to achieve bidirectional alignments using deformable convolutions. To the best of our knowledge, this is the first work to apply deformable convolutions to LF image SR.

III. Network Architecture

In this section, we introduce our LF-DFnet in details. Following [27]–[31], we convert input images from RGB channel space to YCbCr channel space and only super-resolve the Y channel images, leaving Cb and Cr channel images being bicubicly upscaled. Consequently, without considering the channel dimension, an LF can be formulated as a 4D tensor $\mathcal{L} \in \mathbb{R}^{U \times V \times H \times W}$, where U and V represent angular dimensions, H and W represent spatial dimensions. Specifically, a 4D LF can be considered as a $U \times V$ array of SAIs, and the resolution of each SAI is $H \times W$. Following [27]–[31], we achieve LF image SR using SAIs distributed in a square array, i.e., $U = V = A$.

As illustrated in Fig. 1(a), our LF-DFnet takes LR SAIs as its inputs and sequentially performs feature extraction (Section III-A), angular deformable alignment (Section III-B), reconstruction and upsampling (Section III-C).

A. Feature Extraction Module

Discriminative feature representation with rich spatial context information is beneficial to the subsequent feature alignment and SR reconstruction steps. Therefore, a large receptive field with a dense pixel sampling rate is required to extract hierarchical features. To this end, we follow [61] and use residual atrous spatial pyramid pooling (ASPP) module as the feature extraction module in our LF-DFnet.

As shown in Fig. 1(a), input SAIs are first processed by a 1×1 convolution to generate initial features, and then fed to residual ASPP modules (Fig. 1(b)) and residual blocks (Fig. 1(c)) for deep feature extraction. Note that, each view is processed separately and the weights in our feature extraction module are shared among these views. In each residual ASPP block, three 3×3 dilated convolutions (with dilation rates of 1, 2, 4, respectively) are combined in parallel to extract hierarchical features with dense sampling rates. After activation with a Leaky ReLU layer (with a leaky factor of 0.1), features of these three branches are concatenated and fused by a 1×1 convolution. Finally, both the center-view feature $\mathcal{F}_c \in \mathbb{R}^{H \times W \times C}$ and side-view features $\mathcal{F}_i \in \mathbb{R}^{H \times W \times C}$ $(i = 1, 2, \cdots, A^2 - 1)$ are generated by our feature extraction module. Following [28], we set the feature depth to 32 (i.e., $C = 32$). The effectiveness of residual ASPP module is demonstrated in Section V-C.

B. Angular Deformable Alignment Module (ADAM)

Given features generated by the feature extraction module, the main objective of ADAM is to perform alignment between the center-view feature and each side-view feature. Here, we propose a bidirectional alignment approach (i.e., collect-and-distribute) to incorporate angular information. Specifically, side-view features are first aligned with the center-view feature to perform feature collection. Then, these aligned features are fused by a 1×1 convolution to incorporate angular information. Afterwards, the fused features are further aligned with their original features to achieve feature distribution. In this way, angular information can be jointly incorporated into each angular view, and the SR performance of all perspectives can be evenly improved (see Section V-C). In this paper, we cascade K ADAMs to perform feature collection and feature distribution. Without loss of generality, we take the k^{th} $(k = 1, 2, \cdots, K)$ ADAM as an example to introduce its mechanism, as shown in Fig. 1(d).

The core component of ADAM is deformable convolution, which is used to align features according to their corresponding offsets. In our implementation, we use a deformable

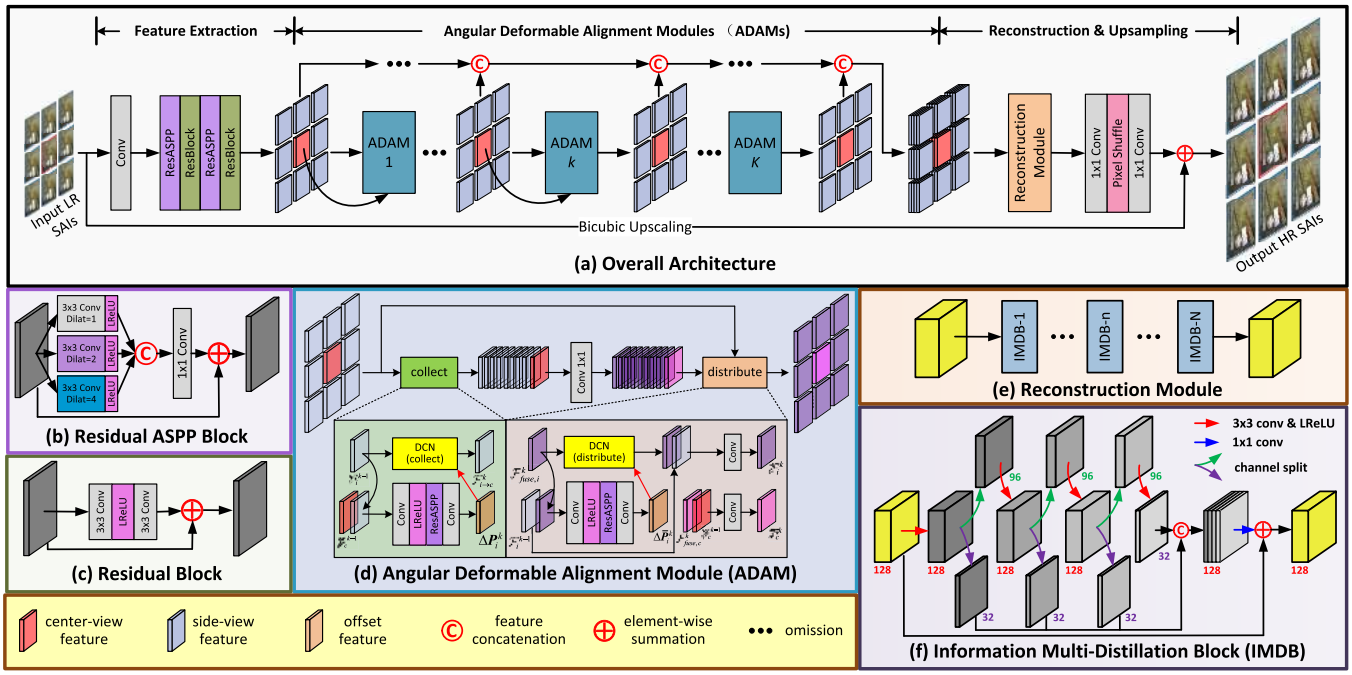

Fig. 1. An overview of our LF-DFnet.

convolution for feature collection and another deformable convolution with shared weights for feature distribution. The first deformable convolution, which is used for feature collection, takes the $(k-1)^{\text{th}}$ side-view feature $\mathcal{F}_i^{k-1}$ and learnable offsets $\Delta \boldsymbol{P}_{i\to c}^k$ as its input to generate $\mathcal{F}_{i\to c}^k$ (which is aligned to the center view). That is,

$$\mathcal{F}_{i\to c}^k = H_{dcn}^k\left(\mathcal{F}_i^{k-1}, \Delta \boldsymbol{P}_{i\to c}^k\right), \tag{1}$$

where H_{dcn}^k represents the deformable convolution in the k^{th} deformable block, $\Delta \boldsymbol{P}_{i\to c}^k = \{\Delta \boldsymbol{p}_n\} \in \mathbb{R}^{H\times W\times C'}$ is the offset of $\mathcal{F}_i^{k-1}$ with respect to $\mathcal{F}_c$. More specifically, for each position $\boldsymbol{p}_0 = (x_0, y_0)$ on $\mathcal{F}_{i\to c}^k$, we have

$$\mathcal{F}_{i\to c}^k(\boldsymbol{p}_0) = \sum_{\boldsymbol{p}_n\in\boldsymbol{R}} w(\boldsymbol{p}_n)\cdot \mathcal{F}_i^{k-1}\left(\boldsymbol{p}_0 + \boldsymbol{p}_n + \Delta \boldsymbol{p}_n\right), \tag{2}$$

where $\boldsymbol{R} = \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$ represents a 3×3 neighborhood region centered at $\boldsymbol{p}_0$. $\boldsymbol{p}_n \in \boldsymbol{R}$ is the predefined integral offset. $\Delta \boldsymbol{p}_n$ is an additional learnable offset, which is added to the predefined offset $\boldsymbol{p}_n$ to make the positions of deformable kernels spatially-variant. Thus, information far away from $\boldsymbol{p}_0$ can be adaptively processed by deformable convolution. Since $\Delta \boldsymbol{p}_n$ can be fractional, we follow [36] to perform bilinear interpolation to generate exact offset values in our implementation.

Since an accurate offset is beneficial to deformable alignment, we design an offset generation branch to learn offset $\Delta \boldsymbol{P}_{i\to c}^k$ in Eq. (1). As illustrated in Fig. 1(d), the side-view feature $\mathcal{F}_i^{k-1}$ is first concatenated with the center-view feature $\mathcal{F}_c$, and then fed to a 1×1 convolution for feature depth reduction. To handle the complicated and large motions between $\mathcal{F}_i^{k-1}$ and $\mathcal{F}_c$, a residual ASPP module (which is identical to that in the feature extraction module but with different weights) is applied to enlarge the receptive field while maintaining a dense sampling rate. The residual ASPP module enhances the exploitation of angular dependencies between the center view and side views, resulting in improved SR performance (see Section V-C). Finally, another 1×1 convolution with $C' = 18$ output channels is used to generate the offset feature.

Once all side-view features are aligned to the center view, a 1×1 convolution is performed to the collected features to fuse the complementary angular information. That is,

$$\mathcal{F}_{fuse}^k = H_{fuse}^k\left(\mathcal{F}_{collect}^k\right), \tag{3}$$

where $\mathcal{F}_{collect}^k = \left[\mathcal{F}_{1\to c}^k, \mathcal{F}_{2\to c}^k, \cdots, \mathcal{F}_{(A^2-1)\to c}^k, \mathcal{F}_c^{k-1}\right]$ denotes the concatenation of the center-view feature and all the aligned side-view features, H_{fuse}^k denotes a 1×1 convolution for angular information incorporation. Finally, $\mathcal{F}_{fuse}^k \in \mathbb{R}^{H\times W\times A^2 C}$ is obtained by performing the above fusion operation.

To super-resolve all LF images, the incorporated angular information in $\mathcal{F}_{fuse}^k$ need to be propagated into each side view. Consequently, we apply deformable convolution for the second time to perform feature distribution. Specifically, the fused feature $\mathcal{F}_{fuse}^k$ is first split along the channel dimension to generate A^2 sub-features. Then, for a specific side view i, a sub-feature $\mathcal{F}_{fuse,i}^k \in \mathbb{R}^{H\times W\times C}$ is concatenated with the corresponding original side-view feature $\mathcal{F}_i^{k-1}$ and fed to the offset generation branch to produce offset $\Delta \bar{\boldsymbol{P}}_i^k$ ($\mathcal{F}_{fuse,i}^k$ with respect to $\mathcal{F}_i^{k-1}$). The final distributed feature of side view i can be obtained according to

$$\mathcal{F}_i^k = H_{squeeze}^k\left(\left[H_{dcn}^k\left(\mathcal{F}_{fuse,i}^k, \Delta \bar{\boldsymbol{P}}_i^k\right), \mathcal{F}_i^{k-1}\right]\right), \tag{4}$$

where $H^k_{squeeze}$ denotes a 1×1 convolution to reduce the channel numbers from $2C$ to C. Note that, the weights of both the offset generation branch and deformable convolution are shared between feature collection and feature distribution. The center-view feature is updated according to

$$\mathcal{F}^k_c = H^k_{squeeze}\left(\left[\mathcal{F}^k_{fuse,c}, \mathcal{F}^{k-1}_c\right]\right). \quad (5)$$

After feature distribution, both the center-view feature $\mathcal{F}^k_c$ and side-view features $\mathcal{F}^k_i, (i = 1, 2, \cdots, A^2 - 1)$ are produced by the k^{th} ADAM. In this paper, we cascade three ADAMs (i.e., $K = 3$) to achieve repetitive feature collection and distribution. Consequently, angular information can be repetitively incorporated into the center view and then propagated to all side views, resulting in notable performance improvements (see Section V-C).

C. Reconstruction & Upsampling Module

To achieve high reconstruction accuracy, the spatial and angular information has to be incorporated. Since preceding modules in our LF-DFnet have produced angular-aligned hierarchical features, a reconstruction module is needed to fuse these features for LF image SR. Following [62], we propose a reconstruction module with information multi-distillation blocks (IMDB). By adopting distillation mechanism to gradually extract and process hierarchical features, superior SR performance can be achieved with a small number of parameters and a low computational cost [63].

The overall architecture of our reconstruction module is illustrated in Fig. 1(e). For each view, the outputs of the feature extraction module and each ADAM are concatenated and fed to several stacked IMDBs. The structure of IMDB is illustrated in Fig. 1(f). Specifically, in each IMDB, the input feature is first processed by a 3×3 convolution and a Leaky ReLU layer. The processed feature is then split into two parts along the channel dimension, resulting in a narrow feature (with 32 channels) and a wide feature (with 96 channels). The narrow feature is preserved and directly fed to the final bottleneck of this IMDB, while the wide feature is fed to a 3×3 convolution to enlarge its channels to 128 for further refinement. In this way, useful information can be gradually distilled, and the SR performance is improved in an efficient manner. Finally, features of different stages in the IMDB are concatenated and processed by a 1×1 convolution for local residual learning.

Features obtained from the reconstruction module are finally fed to an upsampling module. Specifically, a 1×1 convolution is first applied to the reconstructed features to extend their depth to $\alpha^2 C$, where α is the upsampling factor. Then, pixel shuffle is performed to upscale the reconstructed feature to the target resolution $\alpha H \times \alpha W$. Finally, a 1×1 convolution is applied to squeeze the number of feature channels to 1 to generate super-resolved SAIs.

IV. The *NUDT* Dataset

LF images captured by different devices (especially camera arrays) usually have significantly different baseline lengths. It is therefore, necessary to know how existing LF algorithms work under baseline variations, including those developed for depth estimation [73]–[78], view synthesis [79]–[87], and image SR [56], [88]–[91]. However, all existing LF datasets [16], [69]–[72] only include images with fixed baselines. To this end, we introduce a novel LF dataset (namely, the NUDT dataset) with adjustable baselines to facilitate the study of LF algorithms under baseline variations.

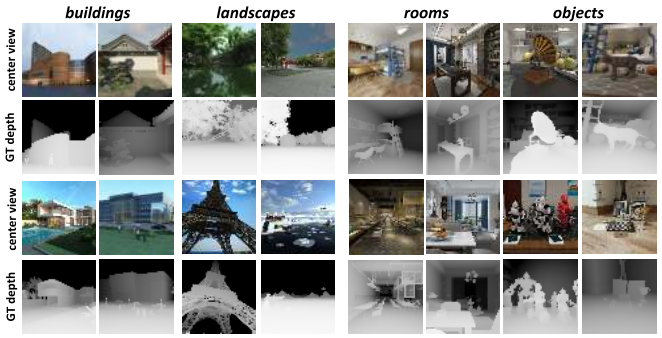

Fig. 2. Example images and their groundtruth depth maps in our NUDT dataset.

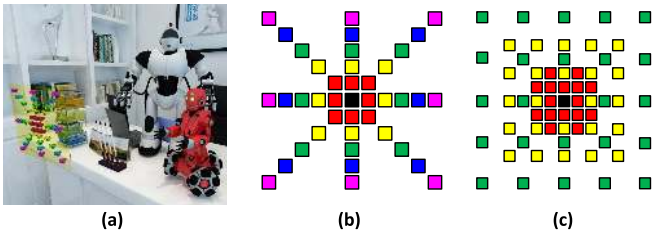

Fig. 3. An illustration of the concentric configuration. (a) Configuration of scene *Robots*. Here, 3×3 camera arrays of 5 different settings of baselines are used as examples. Blocks on the translucent yellow plane denote virtual cameras, where camera arrays of different baselines are drawn in different colors. (b) 3×3 concentric configuration with 5 different settings of baselines. (c) 5×5 concentric configuration with 3 different settings of baselines.

A. Technical Details

Our NUDT dataset has 32 synthetic scenes and covers diverse scenarios (see Fig. 2). All scenes in our dataset are rendered using the 3dsMax software[2], and have an angular resolution of 9×9 and a spatial resolution of 1024×1024. Groundtruth depth maps are available for LF depth/disparity estimation methods. During the image rendering process, all virtual cameras in the array have identical internal parameters and are coplanar with the parallel optical axes. To capture LF images with different baselines, we used a concentric configuration to align camera arrays at the center views. In this way, LF images of different baselines share the same center-view SAI and groundtruth depth map. An illustration of our concentric configuration is shown in Fig. 3. For each scene, we rendered LF images with 5 different baselines. Note that, we tuned the parameters (e.g., lighting, and depth range) to better reflect real scenes. Consequently, our dataset has a

[2]https://www.autodesk.eu/products/3ds-max/overview

TABLE I

MAIN CHARACTERISTICS OF SEVERAL POPULAR LF DATASETS. NOTE THAT, AVERAGE SCORES ARE REPORTED FOR SPATIAL RESOLUTION (SPARES), SINGLE-IMAGE PERCEPTUAL QUALITY METRICS (I.E., BRISQUE [64], NIQE [65], CEIQ [66], ENIQA [67]) AND LF QUALITY ASSESSMENT METRICS (I.E., NRLFQA [68])

Datasets	Type	#Scenes	AngRes	SpaRes (↑)	GT Depth	BRISQUE (↓)	NIQE (↓)	CEIQ (↑)	ENIQA (↓)	NRLFQA (↑)
EPFL [69]	real (lytro)	119	14×14	0.034 Mpx	✗	47.19	5.820	3.286	0.212	2.970
HCInew [70]	synthetic	24	9×9	0.026 Mpx	✓	<u>14.80</u>	3.833	3.153	<u>0.087</u>	2.915
HCIold [71]	synthetic	12	9×9	0.070 Mpx	✓	24.17	**2.985**	<u>3.369</u>	0.117	2.720
INRIA [16]	real (lytro)	57	14×14	0.027 Mpx	✗	23.56	5.338	3.184	0.160	<u>2.983</u>
STFgantry [72]	real (gantry)	12	17×17	**0.118** Mpx	✗	25.28	4.246	2.781	0.232	**2.993**
NUDT (Ours)	synthetic	32	9×9	<u>0.105</u> Mpx	✓	**8.901**	<u>3.593</u>	**3.375**	**0.041**	<u>2.983</u>

Note: 1) Mpx denotes mega-pixels per image. 2) The best results are in **bold** faces and the second best results are <u>underlined</u>. 3) Lower scores of BRISQUE, NIQE, ENIQA and higher scores of CEIQ, NRLFQA indicate better quality.

TABLE II

PUBLIC DATASETS USED IN OUR EXPERIMENTS

Datasets	#Training	#Test
EPFL [69]	70	10
HCInew [70]	20	4
HCIold [71]	10	2
INRIA [16]	35	5
STFgantry [72]	9	2
Total	**144**	**23**

high perceptual quality, which will be introduced in the next subsection.

B. Comparison to Existing Datasets

In this section, we compare our NUDT dataset to several popular LF datasets [16], [69]–[72]. Following [92], we use four no-reference image quality assessment (NRIQA) metrics (i.e, BRISQUE [64], NIQE [65], CEIQ [66], ENIQA [67]) to evaluate the perceptual quality of the center-view images of these datasets. Besides, we also use a no-reference LF quality assessment metric (i.e., NRLFQA [68]) to evaluate the spatial quality and angular consistency of LFs. As shown in Table I, our NUDT dataset achieves the best scores in BRISQUE, CEIQ, and ENIQA, and achieves the second best scores in NIQE and NRLFQA. That is, LF images in our NUDT dataset are angular consistent and have high perceptual quality. Meanwhile, our dataset has more scenes (see #Scenes) and higher image resolution (see SpaRes) than the synthetic HCInew [70] and HCIold [71] datasets.

V. EXPERIMENTS

In this section, we first introduce our implementation details. Then, we compare our LF-DFnet to several state-of-the-art SISR and LF image SR methods from different perspectives. Finally, we present ablation studies to investigate our network.

A. Implementation Details

As listed in Table II, we used 5 public LF datasets in our experiments for both training and test. For the LF datasets (i.e., EPFL [69] and INRIA [16]) recorded by Lytro cameras, we follow [21], [29] to use the *Light Field Toolbox v0.4* [95] to decode raw LF images and extract 4D LF data. All LFs in these datasets have an angular resolution of 9×9. In the training stage, we cropped each SAI into patches with a stride of 32, and used the bicubic downsampling approach to generate LR patches of size 32×32. We performed random horizontal flipping, vertical flipping, and 90-degree rotation to augment the training data by 8 times. Note that, both spatial and angular dimensions need to be flipped or rotated during data augmentation to maintain LF structures.

By default, we used the model with $K = 3$, $N = 4$, $C = 32$, and an angular resolution of 5×5 for both $2\times$ and $4\times$ SR. Our network was trained using the L_1 loss function and optimized using the Adam method [96]. We initialized the weights and bias of the last convolution layer in the offset generation branch with zero values, and used the Kaiming method [97] to initialize other parts of the network. Our LF-DFnet was implemented in PyTorch on a PC with two NVidia RTX 2080Ti GPUs. The batch size was set to 8 and the learning rate was initially set to 2×10^{-4} and decreased by a factor of 0.5 for every 15 epochs. The training was stopped after 50 epochs.

We used PSNR, SSIM, and Perception Index (PI) [93] as quantitative metrics for performance evaluation. Note that, PSNR and SSIM were calculated on the Y channel images and PI was calculated on the RGB channel images. To obtain the metric score (e.g., PSNR) for a dataset with M test scenes (each scene with an angular resolution of $A \times A$), we first calculated this metric on $A \times A$ SAIs on each scene separately, then obtained the score for each scene by averaging its A^2 scores, and finally obtained the score for this dataset by averaging the scores of all M scenes.

B. Comparison to the State-of-the-Arts

We compare our method to several state-of-the-art methods, including 4 single image SR methods (i.e., VDSR [47], EDSR [33], RCAN [51], and ESRGAN [94]) and 5 LF image SR methods (i.e., LFBM5D [24], GB [22], LFSSR [30], resLF [28], and LF-InterNet [31]). For fair comparison, we have retrained all deep-learning methods [28], [30], [31], [33], [47], [51], [94] on the same training datasets as our LF-DFnet. We also include bicubic interpolation as a baseline method. For simplicity, we only present the results on 5×5 LFs for $2\times$ and $4\times$ SR.

1) Quantitative Results: Quantitative results are presented in Table III. Our LF-DFnet achieves the highest PSNR and

TABLE III

PSNR/SSIM/PI VALUES ACHIEVED BY DIFFERENT METHODS FOR 2× AND 4×SR. FOR PSNR AND SSIM, LARGER VALUES INDICATE HIGHER RECONSTRUCTION QUALITY. FOR PI [93], SMALLER VALUES INDICATE HIGHER PERCEPTUAL QUALITY. THE BEST RESULTS ARE IN RED AND THE SECOND BEST RESULTS ARE IN BLUE

Method	Scale	Dataset				
		EPFL [69]	HCInew [70]	HCIold [71]	INRIA [16]	STFgantry [72]
Bicubic	2×	29.50/0.9350/5.633	31.69/0.9335/5.141	37.46/0.9776/5.715	31.10/0.9563/5.592	30.82/0.9473/6.058
VDSR [47]	2×	32.50/0.9599/4.874	34.37/0.9563/4.080	40.61/0.9867/4.211	34.43/0.9742/4.636	35.54/0.9790/4.791
EDSR [33]	2×	33.09/0.9631/4.749	34.83/0.9594/3.914	41.01/0.9875/4.051	34.97/0.9765/4.568	36.29/0.9819/4.642
RCAN [51]	2×	33.16/0.9635/4.780	34.98/0.9602/3.940	41.05/0.9875/4.063	35.01/0.9769/4.591	36.33/0.9825/4.652
LFBM5D [24]	2×	31.15/0.9545/4.965	33.72/0.9548/4.525	39.62/0.9854/4.755	32.85/0.9695/4.998	33.55/0.9718/5.159
GB [22]	2×	31.22/0.9591/4.644	35.25/0.9692/3.741	40.21/0.9879/4.105	32.76/0.9724/4.461	35.44/0.9835/4.469
resLF [28]	2×	32.75/0.9672/4.480	36.07/0.9715/3.678	42.61/0.9922/3.818	34.57/0.9784/4.365	36.89/0.9873/4.580
LFSSR [30]	2×	33.69/0.9748/4.623	36.86/0.9753/3.702	43.75/0.9939/3.755	35.27/0.9834/4.504	38.07/0.9902/4.631
LF-InterNet [31]	2×	34.14/0.9761/4.580	37.28/0.9769/3.658	44.45/0.9945/3.710	35.80/0.9846/4.484	38.72/0.9916/4.602
LF-DFnet (Ours)	2×	34.44/0.9766/4.512	37.44/0.9786/3.623	44.23/0.9943/3.680	36.36/0.9841/4.316	39.61/0.9935/4.549
Bicubic	4×	25.14/0.8311/7.802	27.61/0.8507/7.651	32.42/0.9335/7.644	26.82/0.8860/7.574	25.93/0.8431/7.531
VDSR [47]	4×	27.25/0.8782/6.700	29.31/0.8828/6.417	34.81/0.9518/6.416	29.19/0.9208/6.679	28.51/0.9012/6.503
EDSR [33]	4×	27.84/0.8858/6.293	29.60/0.8874/6.095	35.18/0.9538/6.311	29.66/0.9259/6.248	28.70/0.9075/5.923
RCAN [51]	4×	27.88/0.8863/6.231	29.63/0.8880/5.991	35.20/0.9540/6.233	29.76/0.9273/6.196	28.90/0.9110/5.917
ESRGAN [94]	4×	24.35/0.7968/3.852	26.20/0.8003/3.309	30.69/0.9099/3.317	26.49/0.8652/3.794	25.46/0.8502/4.342
LFBM5D [24]	4×	26.61/0.8689/6.901	29.13/0.8823/6.534	34.23/0.9510/6.579	28.49/0.9137/6.888	28.30/0.9002/6.741
GB [22]	4×	26.02/0.8628/7.217	28.92/0.8842/6.470	33.74/0.9497/6.641	27.73/0.9085/7.220	28.11/0.9014/6.648
resLF [28]	4×	27.46/0.8899/5.509	29.92/0.9011/5.109	36.12/0.9651/5.851	29.64/0.9339/5.615	28.99/0.9214/5.281
LFSSR [30]	4×	28.27/0.9080/5.899	30.72/0.9124/5.504	36.70/0.9690/5.875	30.31/0.9446/5.818	30.15/0.9385/5.815
LF-InterNet [31]	4×	28.67/0.9143/6.061	30.98/0.9165/5.594	37.11/0.9715/5.844	30.64/0.9486/5.919	30.53/0.9426/5.867
LF-DFnet (Ours)	4×	28.77/0.9165/5.836	31.23/0.9196/5.290	37.32/0.9718/5.677	30.83/0.9503/5.649	31.15/0.9494/5.670

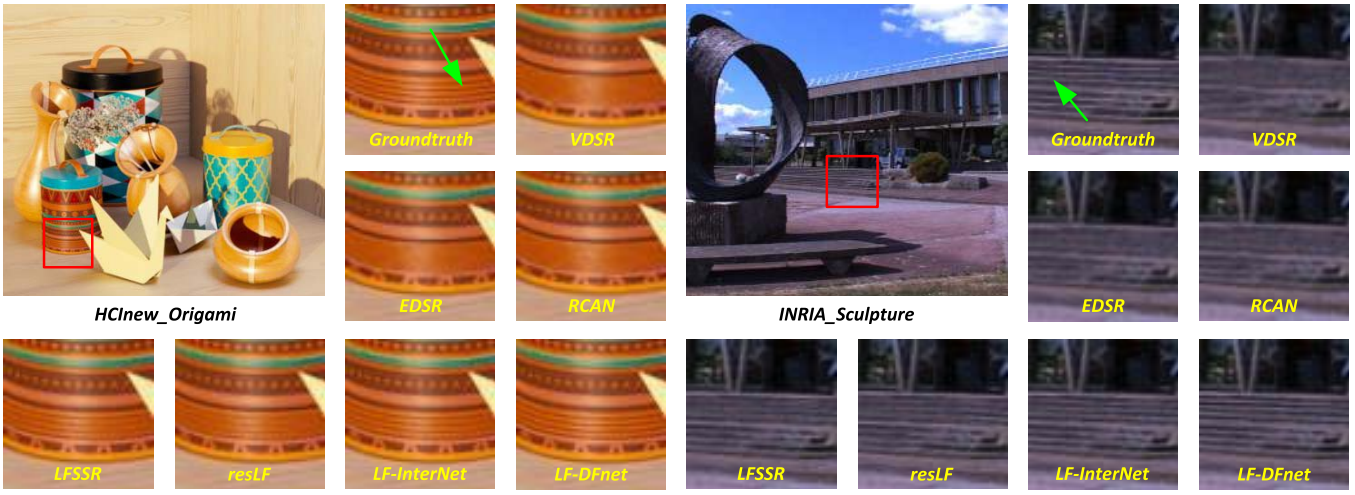

Fig. 4. Visual results of 2×SR.

SSIM scores on all the 5 datasets for 4×SR and on 4 of 5 datasets (i.e., EPFL [69], HCInew [70], INRIA [16] and STFgantry [72]) for 2×SR. In terms of the PI metric [93], our method achieves the state-of-the-art performance for 2×SR, but is slightly inferior to ESRGAN [94] and resLF [28] for 4×SR. Note that, PI is a no-reference metric for perceptual quality evaluation and cannot measure the faithfulness of resultant images. ESRGAN achieves the highest PI scores by generating clear but unfaithful textures (see Fig. 5). It is also worth noting that, the PSNR and SSIM improvements of our LF-DFnet are very significant on the STFgantry dataset for 2×SR. That is because, scenes in the STFgantry dataset are captured by a moving camera mounted on a gantry, and thus have relatively large baselines and significant disparity variations. Our LF-DFnet can handle this disparity problem by using deformable convolution for angular alignment, while maintaining promising performance for LFs with small baselines (e.g., LFs on the EPFL [69] and INRIA [16] datasets). More analyses with respect to different baseline lengths are presented in Section V-B.4.

2) Qualitative Results: Qualitative results for 2× and 4× SR are shown in Figs. 4 and 5, respectively. As compared to the state-of-the-art SISR and LF image SR methods, our method can produce images with more faithful details and less artifacts. Specifically, for 2×SR, the images generated by our LF-DFnet are very close to the groundtruth images. Note that, the stairway in scene *INRIA_Sculpture* and the horizontal stripes in scene *HCInew_Origami* are faithfully recovered by

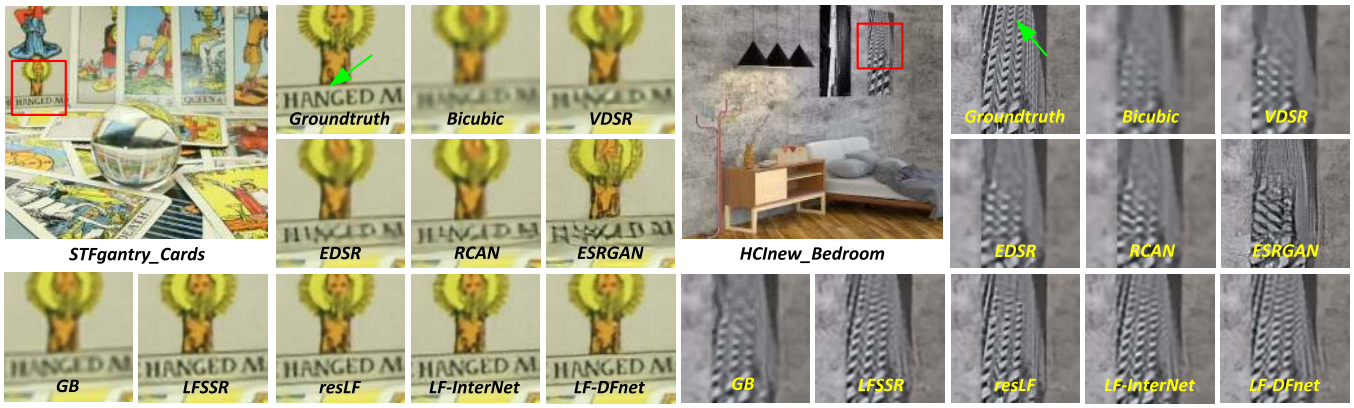

Fig. 5. Visual results of 4×SR.

TABLE IV

COMPARISONS OF THE NUMBER OF PARAMETERS (I.E., #PARAMS.), FLOPS, AND RECONSTRUCTION ACCURACY FOR 2× AND 4× SR. NOTE THAT, FLOPS IS CALCULATED ON AN INPUT LF WITH A SIZE OF 5 × 5 × 32 × 32. HERE, WE USE PSNR AND SSIM VALUES AVERAGED OVER 5 DATASETS [16], [69]–[72] TO REPRESENT THEIR RECONSTRUCTION ACCURACY

Method	Scale	#Params.	FLOPs(G)	PSNR/SSIM
EDSR [33]	2×	38.62M	39.56×25	36.04/0.9737
RCAN [51]	2×	15.31M	15.59×25	36.11/0.9741
resLF [28]	2×	6.35M	37.06	36.57/0.9793
LFSSR [30]	2×	0.81M	25.70	37.53/0.9835
LF-InterNet [31]	2×	4.80M	47.46	38.08/0.9847
LF-DFnet (ours)	2×	3.94M	57.22	38.42/0.9854
EDSR [33]	4×	38.89M	40.66×25	30.20/0.9121
RCAN [51]	4×	15.36M	15.65×25	30.27/0.9133
resLF [28]	4×	6.79M	39.70	30.43/0.9223
LFSSR [30]	4×	1.61M	128.44	31.23/0.9345
LF-InterNet [31]	4×	5.23M	50.10	31.59/0.9387
LF-DFnet (ours)	4×	3.99M	57.31	31.86/0.9415

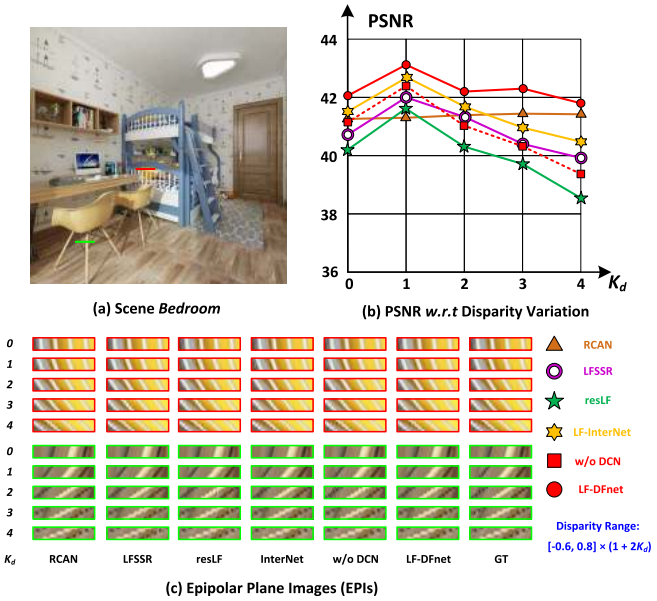

Fig. 6. Performance w.r.t. disparity variations. (a) Scene *Bedroom* of the NUDT dataset. (b) PSNR values achieved by RCAN [51], resLF [28], LFSSR [30], LF-InterNet [31], LF-DFnet, and LF-DFnet without deformable convolution (i.e., w/o DCN) under linearly increased disparities for 2×SR. Our LF-DFnet achieves better SR performance than LFSSR and resLF, especially with large disparity variations (i.e., $K_d \geq 2$), and the deformable convolution contributes to the performance improvements. (c) Epipolar plane images (EPIs) of the corresponding strokes in (a). It can be observed that the misalignment becomes more severe as the baseline length (i.e., K_d) is increased.

our method without blurring or artifacts. For 4×SR, state-of-the-art SISR methods EDSR and RCAN produce blurring results, and the perceptual-oriented SISR method ESRGAN generates images with fake textures. That is because, the SR problem becomes highly ill-posed for 4×SR, and the spatial information in a single image is insufficient to reconstruct high-quality HR images. In contrast, our LF-DFnet can use complementary information among different views to recover missing details, and thus achieves better SR performance.

3) Computational Efficiency: We compare our LF-DFnet to several competitive methods (i.e., EDSR [33], RCAN [51], LFSSR [30], resLF [28], LF-InterNet [31]) in terms of the number of parameters (i.e., #Params) and FLOPs. As shown in Table IV, our method achieves the highest PSNR and SSIM scores with a small number of parameters and FLOPs. Note that, the FLOPs of our method is significantly lower than EDSR and RCAN but slightly higher than resLF and LF-InterNet. That is because, our LF-DFnet uses more complicated feature extraction and reconstruction modules than resLF and LF-InterNet. These modules introduce a notable performance improvement at the cost of a reasonable increase of FLOPs.

4) Performance w.r.t. Disparity Variations: We selected scene *Bedroom* (see Fig. 6(a)) from the NUDT dataset and rendered it with linearly increased baselines ($K_d = 0, 1, \cdots, 4$) to investigate the performance of LF image SR algorithms with respect to disparity variations. Note that, the disparities are proportional to the baseline length when the camera intrinsic parameters (e.g., focal length) are fixed. Following the HCInew dataset [70], we calculated the disparity range (bottom-right in Fig. 6) using the corresponding groundtruth depth values. For performance evaluation, we first used RCAN [51] to evaluate inherent PSNR variation under different baseline lengths, and then compared our LF-DFnet to three state-of-the-art LF image SR methods (i.e., resLF [28], LFSSR [30],

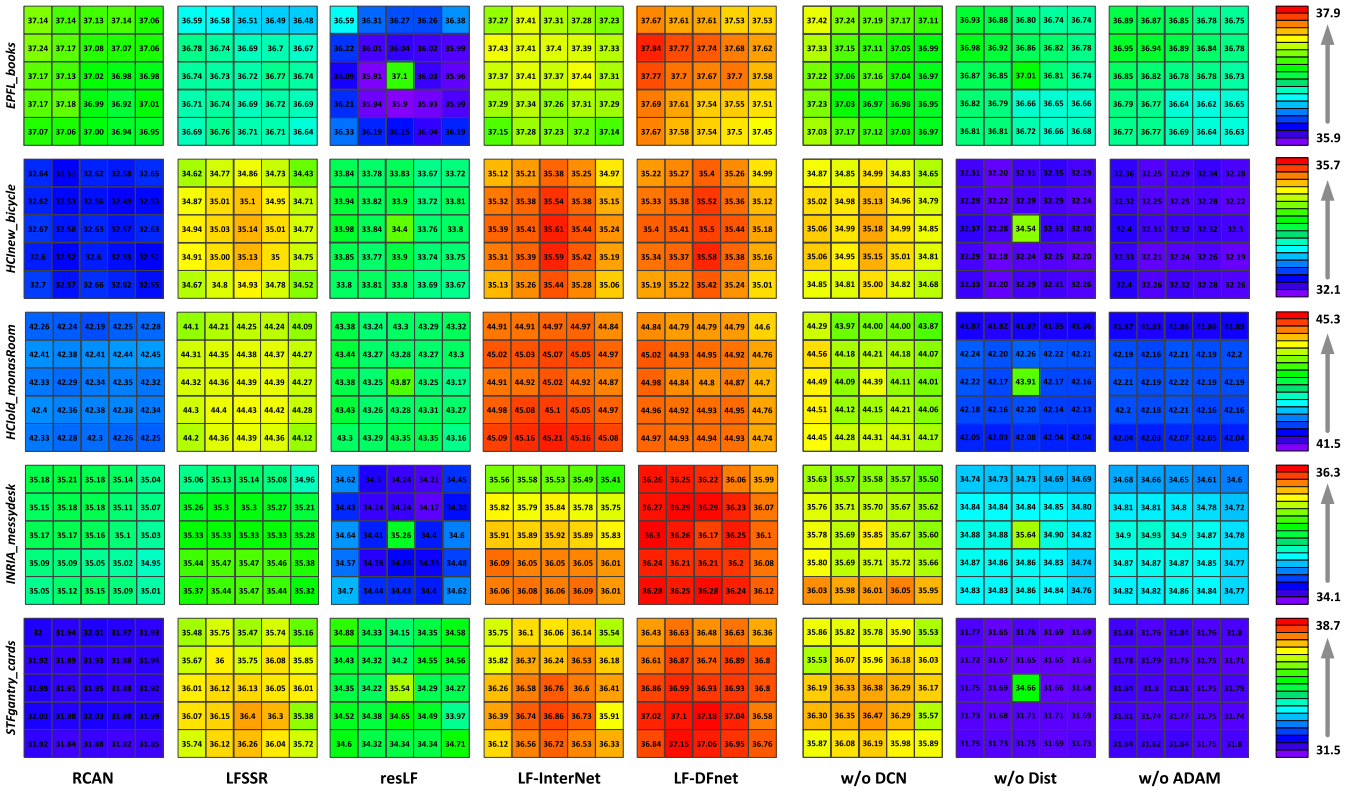

Fig. 7. A visualization of PSNR distribution among different perspectives on 5 × 5 LFs for 2×SR. Here, we compare our LF-DFnet to 4 state-of-the-art SR methods (i.e., RCAN [51], resLF [28], LFSSR [30], LF-InterNet [31]) and three variants of our network (i.e., without using deformable convolution (*w/o DCN*), without performing feature distribution (*w/o Dist*), and without using ADAM (*w/o ADAM*)). Our LF-DFnet achieves high reconstruction quality with a balanced distribution among different perspectives.

and LF-InterNet [31]) and a variant of our network without using deformable convolution (i.e., w/o DCN). Details of this variant are introduced in Section V-C.1. As shown in Fig. 6(b), RCAN achieves comparable SR performance under different baseline settings, which means that the inherent PSNR variation is low. The reconstruction accuracy of LFSSR and resLF drops significantly with increasing disparity variations. In contrast, our LF-DFnet is relatively insensitive to disparity variations and achieves comparable performance to RCAN when $K_d = 4$. Note that, if deformable convolutions are replaced with regular 2D convolutions, our network (i.e., w/o DCN) suffers a notable performance degradation, especially under large baselines (e.g., $K_d > 2$). That is because, large disparities can result in severe misalignment among LF images (see EPIs in Fig. 6(c)) and thus introduce difficulties in angular information exploitation. Since deformable convolution is used to perform feature alignment, our LF-DFnet is more robust to disparity variations, and thus achieves better performance on LF images with wide baselines.

5) Performance w.r.t. Perspectives: Since LF image SR methods aim at super-resolving all SAIs in an LF, we investigate the reconstruction accuracy of different methods with respect to different perspectives. For each dataset listed in Table II, we selected one scene from its test set and calculated the PSNR values on each SAI, as visualized in Fig. 7. We used 5 × 5 central SAIs to perform 2×SR, and compared our LF-DFnet to 4 state-of-the-art SR methods (i.e., RCAN [51], LFSSR [30], resLF [28], LF-InterNet [31]) and 3 variants of our network (i.e., without using deformable convolution (*w/o DCN*), without performing feature distribution (*w/o Dist*), and without using ADAM (*w/o ADAM*)). Details for these variants are introduced in Section V-C.1. As shown in Fig. 7, resLF achieves high PSNR scores on center views but low PSNR scores on side views. That is because, resLF only uses a small part of views to super-resolve side views. The ignored angular information in these discarded views results in the imbalanced PSNR distribution. In contrast, both LF-InterNet and our LF-DFnet achieve improved reconstruction accuracy with a relatively balanced PSNR distribution. It is worth noting that, notable performance drop can be resulted by our LF-DFnet when DCN, ADAM, or feature distribution are removed or canceled. The above experiments clearly demonstrate the effectiveness of our ADAM and collect-and-distribute approach, and illustrate the high reconstruction quality of our LF-DFnet among different perspectives.

6) Performance on Real-World LF Images: We compare our method to RCAN [51], ESRGAN [94], LFSSR [30], resLF [28], and LF-InterNet [31] on real-world LF images by directly applying them to LFs in the EPFL dataset [69]. Since groundtruth HR images are unavailable in this case, we compare the visual performance of different methods. As shown in Fig. 9, our LF-DFnet recovers finer details than RCAN, and produces less artifacts than ESRGAN. It demonstrates that our

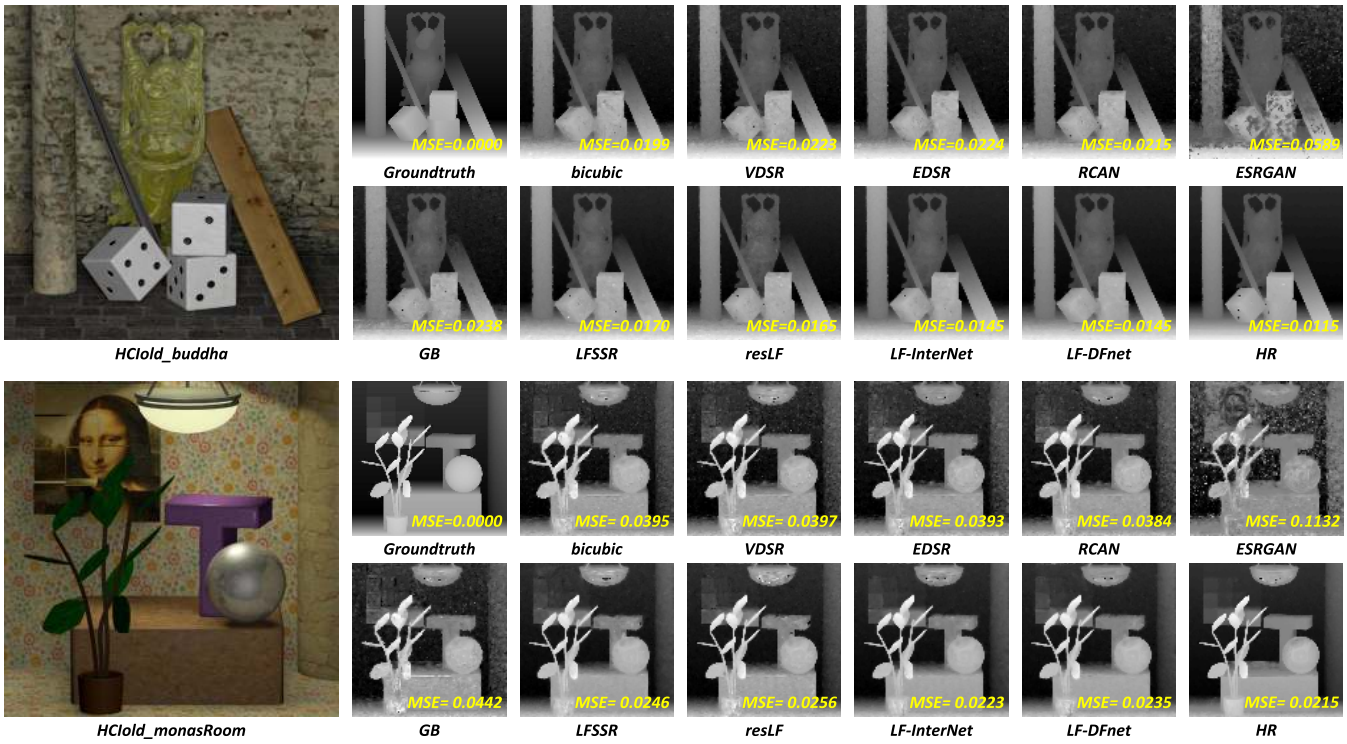

Fig. 8. Depth estimation results achieved by SPO [78] method using 4×SR LF images produced by different SR methods. Note that, the accuracy of depth estimation is improved by using the LF images produced by our LF-DFnet. That is, our LF-DFnet can generate angular-consistent HR LF images which are contributive to depth estimation.

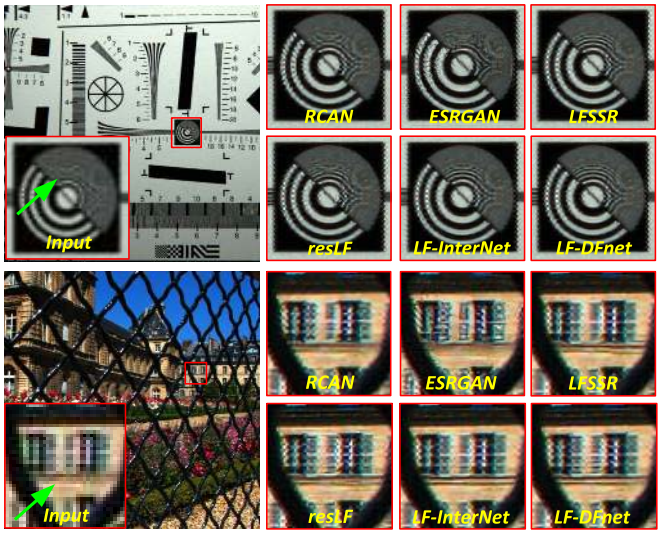

Fig. 9. Visual results achieved by different methods on real-world images.

method can be applied to LF cameras to generate high-quality HR images.

7) Benefits to Depth Estimation: Since high-resolution and angular-consistent LF images are beneficial to depth estimation, we evaluate the reconstruction quality and angular consistency of different SR methods by using their generated LFs to perform depth estimation. Specifically, we used the methods presented in Fig. 5 to perform 4×SR on the HCIold dataset [71], and applied SPO method [78] to the super-resolved LF images to perform depth estimation. Note that, the original HR LFs and LFs generated by bicubic interpolation method were used to produce upper bound results and baseline results, respectively. The mean square error (MSE) metric was used to measure the distance between estimated disparities and groundtruth disparities (normalized to [0, 1]). It can be observed in Fig. 8 that, although SISR methods [33], [47], [51], [94] achieve superior quantitative/visual performance over the bicubic interpolation method (as shown in Table III and Fig. 5), they achieve comparable or even worse results in depth estimation. That is because, SISR methods super-resolve each SAI separately without using any angular information. Consequently, these SISR methods cannot ensure the angular consistency in their generated LFs, which is significantly important to depth estimation. In contrast, several LF image SR methods (i.e., LFSSR, resLF, LF-InterNet, LF-DFnet) can improve the depth estimation performance by generating high-resolution and angular consistent LF images. Note that, the depth estimation results using the output of LF-InterNet and LF-DFnet are very close to the results using the original HR LFs, which clearly demonstrates the high spatial reconstruction quality and angular consistency achieved by these two methods.

C. Ablation Study

In this subsection, we compare our LF-DFnet with several variants to investigate the potential benefits introduced by our network modules and design choices.

TABLE V
PSNR/SSIM VALUES ACHIEVED BY LF-DFNET AND ITS VARIANTS FOR 2×SR

Model	#Params.	Dataset					Average
		EPFL [69]	HCInew [70]	HCIold [71]	INRIA [16]	STFgantry [72]	
LF-DFnet w/o DCN	3.77M	34.04/0.9755	36.94/0.9761	43.63/0.9936	35.93/0.9837	38.73/0.9905	37.85/0.9839 (-0.57/-0.0015)
LF-DFnet w/o ADAM	4.21M	32.81/0.9614	34.73/0.9585	40.94/0.9872	34.70/0.9749	36.40/0.9819	35.92/0.9728 (-2.50/-0.0126)
LF-DFnet w/o Dist	4.07M	32.87/0.9616	34.89/0.9614	41.15/0.9889	34.81/0.9765	36.57/0.9831	36.02/0.9743 (-1.85/-0.0111)
LF-DFnet w/o ASPPinFEM	3.96M	34.32/0.9762	37.21/0.9774	43.89/0.9941	36.24/0.9837	39.37/0.9928	38.21/0.9848 (-0.21/-0.0006)
LF-DFnet w/o ASPPinOFB	4.01M	34.36/0.9764	37.33/0.9781	43.94/0.9941	36.20/0.9838	38.95/0.9923	38.16/0.9849 (-0.26/-0.0005)
LF-DFnet	3.94M	**34.44/0.9766**	**37.44/0.9786**	**44.23/0.9943**	**36.36/0.9841**	**39.61/0.9935**	**38.42/0.9854 (0.00/ 0.0000)**

1) Angular Deformable Alignment Module (ADAM): As the core component of our LF-DFnet, ADAM can perform bidirectional feature alignment between the center view and each side view using deformable convolutions. Here, we validate the effectiveness of ADAM by introducing the following three variants:

- *LF-DFnet w/o DCN*: We replaced the deformable convolution with regular 3×3 convolution in both feature collection and feature distribution stages. Note that, both network depth and feature width of this variant are the same as the original network. Since the offset learning branch was removed with deformable convolution, this variant has a 0.17M parameter reduction as compared to *LF-DFnet*.
- *LF-DFnet w/o ADAM*: We removed all ADAMs in this variant to investigate their contributions. Specifically, we removed both feature collection and feature distribution, and used a residual block (i.e., two 3×3 convolutions and a Leaky ReLU layer) to process each view separately. To achieve fair comparison (i.e., comparable network depth and model size), we increased the number of filters of all convolution layers to make its model size slightly larger than *LF-DFnet*.
- *LF-DFnet w/o Dist*: To investigate the benefit introduced by the bidirectional feature interaction mechanism, we removed feature distribution and only performed feature collection in this variant. Specifically, we used deformable convolution to align side-view features to center view and performed feature fusion as in *LF-DFnet*. However, we did not distribute the incorporated features to side views but followed *LF-DFnet w/o ADAM* to process each side-view feature separately. Similar to *LF-DFnet w/o ADAM*, we adjusted the number of filters to make its model size not smaller than *LF-DFnet*.

Table V shows the comparative results achieved by *LF-DFnet* and its variants. It can be observed in the table that the average PSNR value of *LF-DFnet w/o DCN* suffers a decrease of 0.57 dB as compared to *LF-DFnet*, which demonstrates the effectiveness of deformable convolution in *LF-DFnet*. It is worth noting that, the performance degradation is more significant for the dataset with wide baselines (see *w/o DCN* in Fig. 6 and scores on the STFgantry dataset in Table V). That is because, wide baselines can cause large disparity variations and thus result in severe misalignments among different SAIs. Consequently, the contributive angular information cannot be effectively incorporated without using deformable convolution for feature collection and distribution.

When all ADAMs are removed from *LF-DFnet*, the network (i.e., *LF-DFnet w/o ADAM*) is identical to an SISR model which only uses spatial information within single views to separately super-resolve each SAI. As shown in Table V, *LF-DFnet w/o ADAM* achieves 35.92 dB in average PSNR, which is significantly lower than *LF-DFnet* (38.42 dB) but marginally higher than VDSR [47] (35.49 dB with a 0.66M model). This clearly demonstrates the importance of angular information in LF image SR.

Finally, when feature distribution is canceled, the angular information can be only propagated from side views to center view by *LF-DFnet w/o Dist*, which results in a decrease of 1.85 dB in average PSNR as compared to *LF-DFnet*. It is worth noting that, although angular information is used to super-resolve the center view, as shown in Fig. 7, the PSNR values of *LF-DFnet w/o Dist* on center views are still much lower than those of *LF-DFnet*. That is because, the parameters of the reconstruction module are shared among different perspectives. The unidirectional feature propagation (only from side view to center view) makes the center-view feature significantly vary from the side-view features. This asymmetric feature distribution hinders the reconstruction module to achieve high reconstruction accuracy on both center view and side views.

2) Residual ASPP Module: Residual ASPP module is used in our LF-DFnet for both feature extraction and offset learning. To demonstrate its effectiveness, we introduced two variants (i.e., *LF-DFnet w/o ASPPinFEM* and *LF-DFnet w/o ASPPinOFB*) by replacing the residual ASPP blocks with residual blocks in the feature extraction module and the offset learning branch, respectively. As shown in Table V, *LF-DFnet w/o ASPPinFEM* suffers a 0.21 dB decrease in average PSNR as compared to *LF-DFnet*. That is because, residual ASPP module can extract hierarchical features from input images, which are beneficial to LF image SR. Similarly, a 0.26 dB PSNR decrease is introduced when ASPP module is removed from the offset learning branch. That is because, the ASPP module can achieve accurate offset learning through multi-scale feature representation and the enlargement of receptive fields.

3) Number of ADAMs: We investigate the SR performance with respect to the number of ADAMs in our network. It can be observed in Table VI that the reconstruction accuracy consistently improves as the number of ADAMs increases. However, the improvements tend to be saturated when the

TABLE VI
PSNR/SSIM VALUES ACHIEVED BY LF-DFNET WITH DIFFERENT NUMBER OF ADAMS FOR 2×SR

Model	#Params.	Dataset					Average
		EPFL [69]	HCInew [70]	HCIold [71]	INRIA [16]	STFgantry [72]	
LF-DFnet with 1ADAM	2.52M	33.69/0.9734	36.55/0.9760	43.82/0.9938	35.96/0.9828	39.12/0.9921	37.83/0.9835 (-0.59/-0.0019)
LF-DFnet with 2ADAMs	3.23M	34.40/0.9765	37.38/0.9784	44.06/0.9941	36.33/0.9841	39.40/0.9932	38.31/0.9853 (-0.11/-0.0001)
***LF-DFnet** (3 ADAMs)*	3.94M	34.44/0.9766	37.44/0.9786	44.23/0.9943	36.36/0.9841	39.61/0.9935	38.42/0.9854 (0.00/ 0.0000)
LF-DFnet with 4ADAMs	4.65M	34.47/0.9767	37.51/0.9787	44.34/0.9944	36.35/0.9841	39.76/0.9937	38.49/0.9855 (0.07/ 0.0001)

number of ADAMs is increased from 3 to 4. Since the number of parameters grows linearly with respect to the number of ADAMs, we decided to use 3 ADAMs (i.e., $K = 3$) in our LF-DFnet to achieve a good tradeoff between reconstruction accuracy and computational efficiency.

VI. CONCLUSION AND DISCUSSION

In this paper, we propose an LF-DFnet to achieve LF image SR. Different from existing LF image SR methods, we explicitly handle the disparity problem by performing feature alignment using our designed angular deformable alignment module (ADAM). Moreover, we developed the first baseline-adjustable LF dataset in literature to evaluate the performance of LF image SR methods with respect to disparity variations. Extensive experiments on both public and our self-developed datasets have demonstrated the effectiveness of the proposed method. Our LF-DFnet achieves state-of-the-art SR performance with a small computational cost. The reconstruction accuracy achieved by our LF-DFnet is evenly distributed among angular views and is more robust to disparity variations.

Moreover, as demonstrated in the experiments, our LF-DFnet works well on real-world LF images, and can be used to generate high-quality (i.e., high-resolution and angular-consistent) LFs to benefit downstream tasks (e.g., LF depth estimation). Since high-resolution LF images are needed in many LF tasks (e.g., post-capture refocusing, depth sensing), the experimental results in this paper clearly demonstrates the promising potential applications of our LF-DFnet. In the future, we will try to combine our LF-DFnet with other task-specific networks to further boost the performance gain introduced by high-resolution LF images.

It is worth noting that, although our method achieves promising reconstruction accuracy in terms of PSNR and SSIM values, the visual superiority of our LF-DFnet is very minor and far from satisfactory. That is because, only L_1 loss was used to train our network, and the quality of input LR images are very low, especially for those captured by Lytro cameras. In the future, we will exploit the generative adversarial network (GAN) paradigm to improve the visual quality of reconstructed images and achieve LF image SR with large scaling factors (e.g., 16× SR). We believe that GAN-based LF image SR networks can achieve a better perception-distortion tradeoff by using the additional angular information provided by LF images, and will take a further step toward consumer applications.

REFERENCES

[1] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective light field refocusing for camera arrays using bokeh rendering and superresolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, Jan. 2019.

[2] Z. Xiao, Q. Wang, G. Zhou, and J. Yu, "Aliasing detection and reduction scheme on angularly undersampled light fields," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2103–2115, May 2017.

[3] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu, "PlenoPatch: Patch-based plenoptic image manipulation," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 5, pp. 1561–1573, May 2017.

[4] K. Mishiba, "Fast depth estimation for light field cameras," *IEEE Trans. Image Process.*, vol. 29, pp. 4232–4242, 2020.

[5] A. Chuchvara, A. Barsi, and A. Gotchev, "Fast and accurate depth estimation from sparse light fields," *IEEE Trans. Image Process.*, vol. 29, pp. 2492–2506, 2020.

[6] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, "Unsupervised monocular depth estimation from light field image," *IEEE Trans. Image Process.*, vol. 29, pp. 1606–1617, 2020.

[7] F. Liu, S. Zhou, Y. Wang, G. Hou, Z. Sun, and T. Tan, "Binocular light-field: Imaging theory and occlusion-robust depth perception application," *IEEE Trans. Image Process.*, vol. 29, pp. 1628–1640, 2020.

[8] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5867–5880, Dec. 2019.

[9] H. Sheng, S. Zhang, X. Cao, Y. Fang, and Z. Xiong, "Geometric occlusion analysis in depth estimation using integral guided filter for light-field image," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5758–5771, Dec. 2017.

[10] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu, "Saliency detection via depth-induced cellular automata on light field," *IEEE Trans. Image Process.*, vol. 29, pp. 1879–1889, 2020.

[11] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, "Light field saliency detection with deep convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 4421–4434, 2020.

[12] M. Zhang, J. Li, J. Wei, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 896–906.

[13] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, "Deep learning for light field saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8838–8848.

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

[15] T. Li, D. P. K. Lun, Y.-H. Chan, and Budianto, "Robust reflection removal based on light field imaging," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1798–1812, Apr. 2019.

[16] M. Le Pendu, X. Jiang, and C. Guillemot, "Light field inpainting propagation via low rank matrix completion," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1981–1993, Apr. 2018.

[17] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, and Y. Guo, "DeOccNet: Learning to see through foreground occlusions in light fields," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 118–127.

[18] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

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

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

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

[22] M. Rossi and P. Frossard, "Geometry-consistent light field super-resolution via graph-based regularization," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4207–4218, Sep. 2018.

[23] V. K. Ghassab and N. Bouguila, "Light field super-resolution using edge-preserved graph-based regularization," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1447–1457, Jun. 2020.

[24] M. Alain and A. Smolic, "Light field super-resolution via LFBM5D sparse coding," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2501–2505.

[25] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 24–32.

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

[27] Y. Yuan, Z. Cao, and L. Su, "Light-field image superresolution using a combined deep CNN based on EPI," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1359–1363, Sep. 2018.

[28] S. Zhang, Y. Lin, and H. Sheng, "Residual networks for light field image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11046–11055.

[29] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4274–4286, Sep. 2018.

[30] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, "Light field spatial super-resolution using deep efficient spatial-angular separable convolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2319–2330, May 2019.

[31] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 290–308.

[32] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[33] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[34] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, "Learning for video super-resolution through HR optical flow estimation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Cham, Switzerland: Springer, 2018, pp. 514–529.

[35] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.

[36] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[37] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.

[38] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[39] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019.

[40] H. Wang, D. Su, C. Liu, L. Jin, X. Sun, and X. Peng, "Deformable non-local network for video super-resolution," *IEEE Access*, vol. 7, pp. 177734–177744, 2019.

[41] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, "Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3370–3379.

[42] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," 2020, *arXiv:2004.02803*. [Online]. Available: http://arxiv.org/abs/2004.02803

[43] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–34, Jul. 2020.

[44] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.

[45] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 23, 2020, doi: 10.1109/TPAMI.2020.2982166.

[46] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 184–199.

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

[48] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 114–125.

[49] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[50] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 21, 2020, doi: 10.1109/TPAMI.2020.2968521.

[51] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[52] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 11065–11074.

[53] M. Alain and A. Smolic, "Light field denoising by sparse 5D transform domain collaborative filtering," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2017, pp. 1–6.

[54] K. Egiazarian and V. Katkovnik, "Single image super-resolution via BM3D sparse coding," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 2849–2853.

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

[56] J. Jin, J. Hou, J. Chen, and S. Kwong, "Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2260–2269.

[57] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 331–346.

[58] Y. Zhao, Y. Xiong, and D. Lin, "Trajectory convolution for action recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 2204–2215.

[59] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 529–545.

[60] S. Nah *et al.*, "NTIRE 2019 challenge on video deblurring: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019.

[61] L. Wang *et al.*, "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12250–12259.

[62] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2024–2032.

[63] K. Zhang *et al.*, "AIM 2019 challenge on constrained super-resolution: Methods and results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3565–3574.

[64] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[65] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.

[66] J. Yan, J. Li, and X. Fu, "No-reference quality assessment of contrast-distorted images using contrast enhancement," 2019, *arXiv:1904.08879*. [Online]. Available: http://arxiv.org/abs/1904.08879

[67] X. Chen, Q. Zhang, M. Lin, G. Yang, and C. He, "No-reference color image quality assessment: From entropy to perceptual quality," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, p. 77, Dec. 2019.

[68] L. Shi, W. Zhou, Z. Chen, and J. Zhang, "No-reference light field image quality assessment based on spatial-angular measurement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4114–4128, Nov. 2020.

[69] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2016.

[70] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Cham, Switzerland: Springer, 2016, pp. 19–34.

[71] S. Wanner, S. Meister, and B. Goldlüecke, "Datasets and benchmarks for densely sampled 4D light fields," *Vis., Model. Vis.*, vol. 13, pp. 225–226, Sep. 2013.
[72] V. Vaish and A. Adams, "The (new) Stanford light field archive," Comput. Graph. Lab., Stanford Univ., Tech. Rep., 2008, vol. 6, no. 7. [Online]. Available: http://lightfield.stanford.edu/
[73] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4748–4757.
[74] Williem, I. K. Park, and K. M. Lee, "Robust light field depth estimation using occlusion-noise aware data costs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2484–2497, Oct. 2018.
[75] J. Y. Lee and R.-H. Park, "Complex-valued disparity: Unified depth model of depth from stereo, depth from focus, and depth from defocus based on the light field gradient," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 8, 2019, doi: 10.1109/TPAMI.2019.2946159.
[76] H.-G. Jeon *et al.*, "Depth from a light field image with learning-based matching costs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 297–310, Feb. 2019.
[77] H. Sheng, P. Zhao, S. Zhang, J. Zhang, and D. Yang, "Occlusion-aware depth estimation for light field using multi-orientation EPIs," *Pattern Recognit.*, vol. 74, pp. 587–599, Feb. 2018.
[78] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understand.*, vol. 145, pp. 148–159, Apr. 2016.
[79] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6319–6327.
[80] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on EPI and extended applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1681–1694, Jul. 2019.
[81] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared EPI structure for light field reconstruction," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3261–3273, Jul. 2019.
[82] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 133–147, Jan. 2018.
[83] H. Wing Fung Yeung, J. Hou, J. Chen, Y. Ying Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 137–152.
[84] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, "End-to-end view synthesis for light field imaging with pseudo 4DCNN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 333–348.
[85] S. Zhang, H. Sheng, D. Yang, J. Zhang, and Z. Xiong, "Micro-lens-based matching for scene recovery in lenslet cameras," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1060–1075, Mar. 2018.
[86] N. Meng, H. K.-H. So, X. Sun, and E. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 1, 2019, doi: 10.1109/TPAMI.2019.2945027.
[87] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11141–11148.
[88] N. Meng, X. Wu, J. Liu, and E. Y. Lam, "High-order residual network for light field super-resolution," *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11757–11764.
[89] R. Farrugia and C. Guillemot, "Light field super-resolution using a low-rank prior and deep convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1162–1175, May 2020.
[90] R. A. Farrugia and C. Guillemot, "A simple framework to leverage state-of-the-art single-image super-resolution methods to restore light fields," *Signal Process., Image Commun.*, vol. 80, Feb. 2020, Art. no. 115638.
[91] M. S. K. Gul and B. K. Gunturk, "Spatial and angular resolution enhancement of light fields using convolutional neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2146–2159, May 2018.
[92] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019.
[93] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Proc. Eur. Conf. Comput. Vis. Workshop (ECCVW)*, 2018.
[94] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further improving enhanced super-resolution generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3637–3641.
[95] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.
[96] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[97] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

Yingqian 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 the National University of Defense Technology (NUDT), Changsha, China, in 2018, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology. His research interests focus on low-level vision, particularly on light field imaging and image super-resolution.

Jungang Yang (Member, IEEE) received the B.E. and Ph.D. degrees from the National University of Defense Technology (NUDT), in 2007 and 2013, respectively. He was a visiting Ph.D. student with the University of Edinburgh, Edinburgh, from 2011 to 2012. He is currently an Associate Professor with the College of Electronic Science, NUDT. His research interests include computational imaging, image processing, compressive sensing, and sparse representation. He received the New Scholar Award of the Chinese Ministry of Education in 2012, the Youth Innovation Award, and the Youth Outstanding Talent of NUDT in 2016.

Longguang Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2015, and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2017, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology. His research interests include low-level vision and deep learning.

Xinyi Ying is currently pursuing the M.E. degree with the College of Electronic Science and Technology, National University of Defense Technology (NUDT). Her research interests focus on low-level vision, particularly on image and video super-resolution.

Tianhao Wu received the B.E. degree in electronic engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2020, where he is currently pursuing the M.E. degree with the College of Electronic Science and Technology. His research interests include light field imaging and camera calibration.

Wei An received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 1999. She was a Senior Visiting Scholar with the University of Southampton, Southampton, U.K., in 2016. She is currently a Professor with the College of Electronic Science and Technology, NUDT. She has authored or coauthored over 100 journal and conference publications. Her current research interests include signal processing and image processing.

Yulan Guo (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the National University of Defense Technology (NUDT), in 2008 and 2015, respectively. He was a visiting Ph.D. student with The University of Western Australia from 2011 to 2014. He worked as a Postdoctoral Research Fellow with the Institute of Computing Technology, Chinese Academy of Sciences, from 2016 to 2018. He is currently an Associate Professor. He has authored over 90 articles in journals and conferences, such as the IEEE TPAMI and IJCV. His current research interests focus on 3D vision, particularly on 3D feature learning, 3D modeling, 3D object recognition, and scene understanding. He received the CAAI Outstanding Doctoral Dissertation Award in 2016, and the CAAI Wu-Wenjun Outstanding AI Youth Award in 2019. He served/will serve as an Associate Editor for the *IET Computer Vision* and *IET Image Processing*, a Guest Editor for IEEE TPAMI, an Area Chair for CVPR 2021 and ICPR 2020, an Organizer for a tutorial in CVPR 2016, and a workshop in CVPR 2019.