Light Field Image Super-Resolution Using Deformable Convolution

Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo

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 disparities. All LF images. Moreover, we develop a baseline-adjustable LF dataset to evaluate SR performance under different disparities. Our LF-DFnet can generate high-resolution images with more faithful details and demonstrated the superiority of our method. 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 achieves 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

A THOUGH light field (LF) cameras enable many attractive functions such as post-capture refocusing [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, SAs are first super-resolved using SRCNN [22], 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 [25]–[31].

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 [38], [39], 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 [40], [41] in video SR [42]–[46], 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 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 3dsMax\(^1\) and the baseline of virtual camera arrays is adjustable. In summary, the main contributions of this paper are as follows:

- We propose a LF-DFnet to achieve the state-of-the-art LF image SR performance (as shown in Fig. 1) 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 Section 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 [47]–[49] have been published to comprehensively review SISR methods. Here, we only describe several mile-stone works in literature.

Since Dong et al. [32], [50] 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. [51] proposed a very deep SR network (i.e., VDSR) and achieved a significant performance improvement over SRCNN [32], [50]. Lim et al. [33] proposed an enhanced deep SR network (i.e., EDSR). With the combination of local and global residual connections, EDSR [33] won the NTIRE 2017 SISR challenge [52]. Zhang et al. [53], [54] proposed a residual dense network (i.e., RDN) by introducing a channel attention module and a residual in residual mechanism. More recently, Dai et al. [55] proposed SAN by applying the second-order attention mechanism to SISR. Note that, RCAN [54] and SAN [55] 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 et al. [19] proposed a variational method for LF image SR based on the estimated depth information. Mitra et al. [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 [56] and LF image SR [24] by extending BM3D filtering [57] to LFs. Rossi et al. [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 [50], 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 [33] 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 [58] to LFs. Zhang et al. [28] proposed a multi-stream residual network resLF by stacking SAlS 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. [35] 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 [35] and LF-InterNet [31] are state-of-the-art LF image SR methods to date and can achieve a high reconstruction accuracy.

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 SAlS 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 SAlS 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. [40] 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 [40], [59]–[61], and low-level vision tasks such as video SR [42]–[45]. Specifically, Tian et al. [42] 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. [43] 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 [43] won the NTIRE19 video restoration and enhancement challenges [62]. More recently, deformable convolution is integrated with non-local operation [44], convolutional LSTM [45] and 3D convolutions [46] to further enhance the video SR performance.

In summary, existing deformable convolution-based video SR methods [42]–[46] 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 \( L \in \mathbb{R}^{U \times V \times H \times W} \), 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 SAlS, and the resolution of each SAI is \( H \times W \). Following [27]–[31], we achieve LF image SR using SAlS distributed in a square array, i.e., \( U = V = A \).

As illustrated in Fig. 2(a), our LF-DFnet takes LR SAlS 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 [63] and use residual atrous spatial pyramid pooling (ASPP) module as the feature extraction module in our LF-DFnet.

As shown in Fig. 2(a), input SAlS are first processed by a \( 1 \times 1 \) convolution to generate initial features, and then fed to residual ASPP modules (Fig. 2(b)) and residual blocks (Fig. 2(c)) for deep feature extraction. Note that, each view are processed separately and the weights in our feature extraction module are shared among these views. In each residual ASPP block, three \( 3 \times 3 \) 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 \( 1 \times 1 \) convolution. Finally, both the center-view feature \( F_c \in \mathbb{R}^{H \times W \times C} \) and side-view features \( F_i \in \mathbb{R}^{H \times W \times C} (i = 1, 2, \ldots, A^2 - 1) \) 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 warped to the center view and aligned with the center-view feature (i.e., feature collection). These aligned features are fused by a $1 \times 1$ convolution to incorporate angular information. Afterwards, the fused feature is warped to side views by performing alignment with their original features (i.e., 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. In this paper, we cascaded $K$ 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. 2(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 convolution for feature collection and another deformable convolution with opposite offset values for feature distribution. The first deformable convolution, which is used for feature collection, takes the $(k-1)^{th}$ side-view feature $F_{i-1}^{k-1}$ and learnable offsets $\Delta P^{k}_{i}$ as its input to generate the $k^{th}$ feature $F_{i-1}^{k}$ (which is aligned to the center view). That is,

$$F_{i-1}^{k} = H_{dcn}^{k} (F_{i-1}^{k-1}, \Delta P_{i}^{k}),$$

where $H_{dcn}^{k}$ represents the deformable convolution in the $k^{th}$ deformable block, $\Delta P_{i}^{k} = \{ \Delta p_{n} \} \in \mathbb{R}^{H \times W \times C'}$ is the offset of $F_{i-1}^{k-1}$ with respect to $F_{c}$. More specifically, for each position $p_{0} = (x_{0}, y_{0})$ on $F_{i-1}^{k}$, we have

$$F_{i-1}^{k}(p_{0}) = \sum_{p_{n} \in R} w(p_{n}) \cdot F_{i-1}^{k-1}(p_{0} + p_{n} + \Delta p_{n}),$$

where $R = \{(-1, -1), (-1, 0), \cdots, (0, 1), (1, 1)\}$ represents a $3 \times 3$ neighborhood region centered at $p_{0}$. $p_{n} \in R$ is the predefined integral offset. $\Delta p_{n}$ is an additional learnable offset, which is added to the predefined offset $p_{n}$ to make the positions of deformable kernels spatially-variant. Thus, information far away from $p_{0}$ can be adaptively processed by deformable convolution. Since $\Delta p_{c}$ can be fractional, we follow [40] to perform bilinear interpolation in our implementation.

Since an accurate offset is beneficial to deformable alignment, we design an offset generation branch to learn offset $\Delta P_{i}^{k}$ in Eq. (1). As illustrated in Fig. 2(d), the side-view feature $F_{i}^{k-1}$ is first concatenated with the center-view feature $F_{c}$, and then fed to a $1 \times 1$ convolution for feature depth reduction. To handle the complicated and large motions between $F_{i}^{k-1}$ and $F_{c}$, a residual ASPP module (which is identical to that in Section IV-A) 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. The effectiveness of the residual ASPP module in the offset generation branch is investigated in Section V-C. Finally, another $1 \times 1$ 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 $1 \times 1$ convolution is performed to fuse the angular information in these aligned features.

$$F_{c}^{k} = H_{1 \times 1}^{k} \left( [F_{1-1,c}^{k}, F_{2-1,c}^{k}, \cdots, F_{K-1-1,c}^{k} \mid F_{c}] \right),$$

where $[\cdot, \cdot]$ denotes concatenation and $H_{1 \times 1}^{k}$ denotes a $1 \times 1$ convolution.

To super-resolve all LF images, the incorporated angular information need to be encoded into each side view. Con-
Fig. 3: Example images and their groundtruth depth maps in our NUDT dataset.

sequently, we perform feature distribution to propagate the incorporated angular information to side views. Since the disparities between side-view features and center-view features are mutual, we do not perform additional offset learning. Instead, we use the opposite offset $\Delta P^k = -\Delta P^k$ to warp the fused center-view feature $F^k_c$ to the $i^{th}$ side view. That is,

$$F^k_c = H_{den} (F^k_c, \Delta P^k_i).$$ \hspace{1cm} (4)

After feature distribution, both the center-view feature $F^k_c$ and side-view features $F^k_i, (i = 1, 2, \cdots, A^2-1)$ are produced by the $k^{th}$ ADAM. In this paper, we cascade four ADAMs 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 continuous 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 [64], 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 [65].

The overall architecture of our reconstruction module is illustrated in Fig. 4(e). For each view, the outputs of the feature extraction module and each ADAM are concatenated and processed by a $1 \times 1$ convolution for coarse fusion. Then, the coarsely-fused feature (with 128 channels) is fed to several stacked IMDBs for deep feature fusion. The structure of IMDB is illustrated in Fig. 4(f). Specifically, in each IMDB, the input feature is first processed by a $3 \times 3$ convolution and a Leaky ReLU. 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 $3 \times 3$ 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 $1 \times 1$ convolution for local residual learning. Moreover, the feature produced by the last IMDB is processed by a $3 \times 3$ convolution to reduce its depth from 128 to 32 for global residual learning.

Features obtained from the reconstruction module are finally fed to a upsampling module. Specifically, a $1 \times 1$ convolution is first applied to the reconstructed features to extend their depth to $\alpha^2 C,$ where $\alpha$ 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 $1 \times 1$ 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 [72]–[77], view synthesis [78]–[86], and image SR [35], [87]–[90]. However, all existing LF datasets [16], [36], [57], [70], [71] only include images with fixed baselines. To facilitate the study of LF algorithms under baseline variations, we introduce a novel LF dataset (namely, the NUDT dataset) with adjustable baselines, which is available at: https://github.com/YingqianWang/NUDT-Dataset.

A. Technical Details

Our NUDT dataset has 32 synthetic scenes and covers diverse scenarios (see Fig. 5). All scenes in our dataset are rendered using the 3dsMax software\footnote{https://www.autodesk.com/products/3ds-max/overview} and have an angular resolution of $9 \times 9$ and a spatial resolution of $1024 \times 1024$. 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. 4. For each scene, we rendered LF images with 10 different baselines. Note that, we tuned the parameters (e.g., lighting, and depth range) to better reflect real scenes. Consequently, our dataset has a 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, 36, 37, 70, 71]. Following [91], we use four no-reference image quality assessment (NRIQA) metrics to evaluate the perceptual quality of LF images in these datasets. These NRIQA metrics, including blind/referenceless image spatial quality evaluator (BRISQUE) [66], natural image quality evaluator (NIQE) [67], contrast enhancement based image quality evaluator (CEIQ) [68], and entropy-based image quality assessment (ENIQA) [69], are highly correlated to human perception. As shown in Table I, our NUDT dataset achieves the best scores in BRISQUE [66], CEIQ [68], and ENIQA [69], and achieves the second best score in NIQE [67]. That is, images in our NUDT dataset have high perceptual quality. Meanwhile, our dataset has more scenes (see #Scenes) and higher image resolution (see SpaRes) than the HCInew [36] and the HCInd [71] datasets.

V. EXPERIMENTS

In this section, we first introduce our implementation details. Then, we compare our LF-DFnet to state-of-the-art SISR and LF image SR methods. 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. All LFs in these datasets have an angular resolution of $9 \times 9$. In the training stage, we cropped each SAI into HR patches with a stride of 32, and used the bicubic downsampling approach to generate LR patches with a resolution of $64 \times 64$. 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 $K = 4$, $N = 4$, $C = 32$, and an angular resolution of $5 \times 5$ for both $2 \times$ and $4 \times$ SR. We also investigated several variants of our LF-DFnet in Section V.C. The $L_1$ loss function was used to train our network due to its robustness to outliers [94].

Following [28], [29], [30], [31], [32], [33], we used PSNR and SSIM as quantitative metrics for performance evaluation. Both PSNR and SSIM were separately calculated on the Y channel of each SAI. To obtain the overall metric score for a dataset with $M$ test scenes (each scene with an angular resolution of $A \times A$), we first obtained the score for a scene by averaging its $A^2$ scores, and then generated the overall score by averaging the scores of all $M$ scenes.

Our LF-DFnet was implemented in PyTorch on a PC with two NVidia RTX GPUs. Our model was initialized using the Xavier method [96] and optimized using the Adam method [97]. The batch size was set to 8 and the learning rate was initially set to $4 \times 10^{-4}$ and decreased by a factor of 0.5 for every 10 epochs. The training was stopped after 50 epochs and took about 1.5 days.

### B. Comparison to the State-of-the-arts

We compare our method to several state-of-the-art methods, including 6 single image SR methods (i.e., VDSR [51], EDSR [53], RCAN [54], SAN [55], SRGAN [92], and ESRGAN [93]) and 7 LF image SR methods (i.e., LFBM5D [24], GB [22], LFNet [29], LFSSR [50], resLF [28], LF-ATO [55], and LF-InterNet [31]). We also use bicubic interpolation method to present baseline results.

1) Quantitative Results: Quantitative results are presented in Table III. Our LF-DFnet achieves the highest SSIM scores on all the 5 datasets for both $2 \times$ and $4 \times$ SR. In terms of PSNR, our method achieves the best performance on the HCInew and
TABLE III: PSNR/SSIM values achieved by different methods for $2 \times$ and $4 \times$SR. The best results are in red and the second best results are in blue.

| Method             | Scale | Dataset            | EPFL [19] | HCInet [20] | HCIold [18] | INRIA [17] | STFgantry [17] |
|--------------------|-------|--------------------|-----------|-------------|-------------|-------------|----------------|
| Bicubic            | $2 \times$ | 29.50 / 0.955 | 31.69 / 0.944 | 37.46 / 0.978 | 31.10 / 0.986 | 30.82 / 0.974 |
| VDSR [51]          | $2 \times$ | 32.01 / 0.959 | 34.37 / 0.956 | 40.34 / 0.986 | 33.80 / 0.972 | 35.80 / 0.980 |
| EDSSR [35]         | $2 \times$ | 32.86 / 0.965 | 35.02 / 0.961 | 41.11 / 0.988 | 36.41 / 0.977 | 37.08 / 0.985 |
| RCAN [35]          | $2 \times$ | 33.46 / 0.967 | 35.56 / 0.963 | 41.59 / 0.989 | 35.18 / 0.978 | 38.18 / 0.988 |
| SAN [55]           | $2 \times$ | 33.36 / 0.967 | 35.51 / 0.963 | 41.47 / 0.989 | 35.15 / 0.978 | 37.98 / 0.987 |
| LFBSRD [24]        | $2 \times$ | 31.15 / 0.956 | 33.72 / 0.955 | 39.62 / 0.985 | 32.85 / 0.969 | 33.55 / 0.972 |
| GB [22]            | $2 \times$ | 31.22 / 0.959 | 35.25 / 0.969 | 40.21 / 0.988 | 32.76 / 0.972 | 35.44 / 0.984 |
| LFNet 29 [20]      | $2 \times$ | 31.79 / 0.950 | 33.52 / 0.943 | 39.44 / 0.982 | 33.49 / 0.966 | 32.76 / 0.957 |
| LFSSR [50]         | $2 \times$ | 34.15 / 0.973 | 36.98 / 0.974 | 43.29 / 0.993 | 35.76 / 0.982 | 37.67 / 0.989 |
| resLF [28]         | $2 \times$ | 33.22 / 0.969 | 35.79 / 0.969 | 42.30 / 0.991 | 34.86 / 0.979 | 36.28 / 0.985 |
| LF-ATO [35]        | $2 \times$ | 34.49 / 0.976 | 37.28 / 0.977 | 43.76 / 0.994 | 36.21 / 0.984 | 39.06 / 0.992 |
| LF-InterNet [31]   | $2 \times$ | 34.76 / 0.976 | 37.20 / 0.976 | 44.65 / 0.995 | 36.64 / 0.984 | 38.48 / 0.991 |
| LF-DFnet (Ours)    | $2 \times$ | 34.37 / 0.977 | 37.77 / 0.979 | 44.64 / 0.995 | 36.17 / 0.985 | 40.17 / 0.994 |

STFgantry datasets for $2 \times$ and $4 \times$SR, and on the HCIold dataset for $4 \times$SR. On datasets captured by Lytro cameras (i.e., EPFL and INRIA), our method is marginally inferior to LF-ATO and LF-InterNet but significantly better than other methods (e.g., RCAN, SAN, and resLF). It is worth noting that, the superiority of our LF-DFnet is very significant on the STFgantry dataset for $2 \times$ 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 convolutions for angular alignment, while maintaining promising performance for LFs with small baselines (e.g., LFs on the EPFL and INRIA datasets). More analyses with respect to different baseline lengths are presented in Section V-B5.

2) Qualitative Results: Qualitative results for $2 \times$ and $4 \times$ SR are shown in Figs. 5 and 6 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 $2 \times$ SR, the images generated by our LF-DFnet are very close to the groundtruth images. Note that, the stairway in scene INRIA_Sculpture is faithfully recovered by our method without blurring or artifacts, and the tile edges in the scene HCIold_buddha are as sharp as in the groundtruth image. For $4 \times$ SR, state-of-the-art SISR methods RCAN and SAN produce blurring results with warped textures, and the perceptual-oriented SISR method ESRGAN generates images with fake textures. That is because, the SR problem becomes highly ill-posed for $4 \times$ 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 superior SR performance.

3) Computational Efficiency: We compare our LF-DFnet to several competitive methods [28], [31], [34], [35], [55] 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 are significantly lower than RCAN, SAN, and LF-ATO but marginally 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

TABLE IV: Comparisons of the number of parameters (i.e., #Params) and FLOPs for $2 \times$ and $4 \times$ SR. Note that, FLOPs is calculated on an input LF with a size of $5 \times 5 \times 32 \times 32$. Here, we use PSNR and SSIM values averaged over 5 datasets [16], [56], [57], [70], [71] to represent their reconstruction accuracy.

| Method     | Scale | #Params. | FLOPs(G) | PSNR / SSIM |
|------------|-------|----------|----------|-------------|
| RCAN [35]  | $2 \times$ | 15.44M | 15.71M | 36.79 / 0.977 |
| SAN [55]   | $2 \times$ | 15.71M | 16.05M | 36.69 / 0.979 |
| resLF [28] | $2 \times$ | 6.35M  | 37.06    | 34.49 / 0.979 |
| LF-ATO [35] | $2 \times$ | 1.51M | 57.96    | 38.16 / 0.985 |
| LF-InterNet [31] | $2 \times$ | 4.80M | 47.46    | 38.25 / 0.985 |
| LF-DFnet   | $2 \times$ | 4.62M  | 57.21    | 38.62 / 0.986 |
| RCAN [35]  | $4 \times$ | 15.59M | 16.34M | 31.01 / 0.925 |
| SAN [55]   | $4 \times$ | 15.86M | 16.67M | 31.00 / 0.925 |
| resLF [28] | $4 \times$ | 6.79M  | 39.70    | 30.74 / 0.927 |
| LF-ATO [35] | $4 \times$ | 1.66M | 686.99   | 31.89 / 0.940 |
| LF-InterNet [31] | $4 \times$ | 5.23M | 50.31    | 31.97 / 0.939 |
| LF-DFnet   | $4 \times$ | 4.64M  | 59.85    | 32.00 / 0.943 |
increase of FLOPs.

4) Performance w.r.t. Perspectives: Since LF image SR methods aim at super-resolving all SAIs in an LF, we compare our method to resLF under different perspectives. We used the central 3×3, 5×5, 7×7, and 9×9 SAIs in the HCInew dataset to perform 2×SR, and used PSNR values for performance evaluation, as visualized in Fig. 7. Note that, due to the changing perspectives, the contents of different SAIs are not identical, resulting in inherent PSNR variations among perspectives. Therefore, we evaluate this variation by using RCAN to perform SISR on each SAI. As shown in Fig. 7, RCAN achieves a relatively balanced PSNR distribution (Std=0.0327 for 9×9 LFs). It demonstrates that the inherent PSNR variation among perspectives are relatively small. It can be observed that resLF achieves notable performance improvements over RCAN under all angular resolutions. However, since resLF uses part of views for LF image SR, the PSNR scores achieved by resLF on side views are relatively low.

As compared to RCAN and resLF, our method uses all SAIs to super-resolve each view and handles the disparity problem. Consequently, our method can achieve better SR performance (i.e., higher PSNR values) with a more balanced distribution (i.e., lower Std scores). It can be also observed that, our SR performance is continuously improved as the
angular resolution increases from 3×3 to 7×7. That is because, the additional views can introduce more angular information, which is beneficial to SR reconstruction. Note that, our LF-DFnet achieves comparable performance with 7×7 and 9×9 input views (37.91 vs. 37.89 in average PSNR score). That is, the angular information tends to be saturated when angular resolution is more than 7×7, and a further increase in angular resolution cannot introduce significant performance improvement.

5) Performance w.r.t. Disparity Variations: We selected 4 scenes (see Fig. 8) from our NUDT dataset, and rendered them with linearly increased baselines. Note that, the disparities in a specific scene are proportional to the baseline length when the camera intrinsic parameters (e.g., focal length) are fixed. Consequently, we can investigate the performance of LF image SR algorithms with respect to disparity variations by straightforwardly applying them to the same scenes rendered with different baselines. Following the HCInew dataset [36], we calculated the disparity range of each scene using its groundtruth depth value. As shown in Fig. 8, the reconstruction accuracy (i.e., PSNR) of both methods tends to be decreased with increasing disparities, except for LF-DFnet on scene Robots (see Fig. 8(c)). Note that, the superiority of our LF-DFnet becomes more significant on LF images with larger disparity variations (i.e., wider baselines). That is because, large disparities can result in large misalignment among LF images and thus introduce difficulties in angular information exploitation. Since deformable convolution is used by our method to perform angular alignment, our LF-DFnet is more robust to disparity variations, and thus achieves better performance on LF images with wide baselines.

6) Performance Under Real-World Degradations: We compare our method to RCAN, ESRGAN, resLF, LF-ATO, and LF-InterNet under real-world degradation by directly applying them to LFs in the EPFL dataset. Since no groundtruth HR images are available in this dataset, we compare their visual performance in Fig. 9. Our LF-DFnet recovers much finer details from input LF images, and produces less artifacts than RCAN and ESRGAN. Since LF structures keep unchanged between bicubic and real-world degradation, our method can successfully learn to incorporate spatial and angular information from bicubically downsampled training data, and is well generalized to LF images under real degradation.

C. Ablation Study

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

1) ADAMs: As the core component of our LF-DFnet, ADAM can perform feature alignment between the center...
Fig. 9: Visual results achieved by different methods under real-world degradation.

view and each side view. Here, we investigate ADAM by introducing the following three variants:

- **RegularCNN**: it is introduced by replacing the deformable convolutions with regular $3 \times 3$ convolutions in both feature collection and feature distribution stages.
- **RegularDist**: it is introduced by replacing the deformable convolutions with regular $3 \times 3$ convolutions only in feature distribution stage.
- **LearnDist**: it is introduced by performing offset learning during feature distribution rather than using their opposite values.

Apart from these three variants, SR performance is also influenced by the number of ADAMs in the network. To investigate the effect of these coupled factors, we trained 20 models with 4 design options and 5 different numbers of ADAMs from scratch. Comparative results of these 20 models are shown in Fig. 10.

It can be observed from Figs. 10(a) and 10(d) that RegularDist, LearnDist, and our model achieve comparable results in center-view PSNR/SSIM, while RegularCNN achieves relatively lower scores. That is because, by using deformable convolutions for feature collection, contributive information can be effectively collected and used to reconstruct the center views. Similarly, deformable convolutions in the feature distribution stage also play an important role in LF image SR. As shown in Figs. 10(c) and 10(f), RegularCNN and RegularDist achieve lower PSNR/SSIM scores than LearnDist and the proposed model. That is because, without deformable convolutions for feature distribution, the incorporated information cannot be effectively fused to side views, resulting in a lower minimum reconstruction accuracy. In terms of averaged PSNR/SSIM (see Figs. 10(b) and 10(e)), the proposed model and LearnDist achieve comparable results. However, LearnDist performs offset learning twice in each ADAM, and thus has larger model size and higher FLOPs than the proposed model, as shown in Figs. 10(g) and 10(h). In summary, our proposed model can achieve good SR performance on all angular views while maintaining a reasonable computational cost.

Moreover, it can be observed in Figs. 10(a), 10(f) that the reconstruction accuracy is improved as the number of ADAMs increases. However, the performance tends to be saturated when the number of ADAM is increased from 4 to 5. Since the model size (i.e., Params.) and memory cost (i.e., FLOPs) grow linearly with respect to the number of ADAMs (as shown in Figs. 10-g and 10-h), we finally use 4 ADAMs in our network (i.e., $K = 4$) to achieve a tradeoff between reconstruction accuracy and computational efficiency.

TABLE V: Average PSNR and SSIM values achieved on 5 datasets [16], [36], [37], [70], [71] by LF-DFnet and its variants for 2× SR. Here, the number of parameters (i.e., #Params.) and FLOPs (calculated with an input SAI of size 160 × 160) are reported to show their computational efficiency.

| Model                  | PSNR | SSIM | #Params. | FLOPs |
|------------------------|------|------|----------|-------|
| LF-DFnet_woASPPinFEM   | 38.46| 0.9856| 4.60M    | 56.6G |
| LF-DFnet_woASPPinFES   | 38.44| 0.9856| 4.57M    | 56.1G |
| LF-DFnet               | 38.59| 0.9858| 4.62M    | 57.2G |

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 LF-DFnet_woASPPinFEM and LF-DFnet_woASPPinFES 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_woASPPinFEM suffers a 0.13 dB decrease in 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 0.16 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.

VI. CONCLUSION

In this paper, we proposed an LF-DFNet to handle the disparity problem for LF image SR. By performing feature alignment using our angular deformable alignment module, the angular information can be well incorporated and the SR performance is significantly improved. Moreover, we develop a baseline-adjustable LF dataset for performance evaluation. Experimental results on both public and our self-developed datasets have demonstrated the superiority of our method. Our LF-DFNet achieves state-of-the-art quantitative and qualitative SR performance, and is more robust to disparity variations.

REFERENCES

[1] C.-T. Huang, Y.-W. Wang, L.-R. Huang, J. Chin, and L.-G. Chen, “Fast physically correct refocusing for sparse light fields using block-based multi-rate view interpolation,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 603–618, 2016.

[2] Z. Xiao, Q. Wang, G. Zhou, and J. Yu, “Aliasing detection and reduction scheme on angularly undersampled light fields,” IEEE Transactions on Image Processing, vol. 26, no. 5, pp. 2103–2115, 2017.

[3] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, “Selective light field re-focusing for camera arrays using bokeh rendering and super-resolution,” IEEE Signal Processing Letters, vol. 26, no. 1, pp. 204–208, 2018.

[4] K. Mishiba, “Fast depth estimation for light field cameras,” IEEE Transactions on Image Processing, vol. 29, pp. 4232–4242, 2020.

[5] A. Chuchvara, A. Barsi, and A. Gotchev, “Fast and accurate depth estimation from sparse light fields,” IEEE Transactions on Image Processing, vol. 29, pp. 1606–1617, 2019.

[6] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, “Unsupervised monocular depth estimation from light field image,” IEEE Transactions on Image Processing, vol. 29, pp. 1606–1617, 2019.

[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 Transactions on Image Processing, vol. 29, pp. 1628–1640, 2019.

[8] J. Shi, X. Jiang, and C. Gullemot, “A framework for learning depth from a flexible subset of dense and sparse light field views,” IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5867–5880, 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 Transactions on Image Processing, vol. 26, no. 12, pp. 5758–5771, 2017.

[10] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu, “Saliency detection via depth-induced cellular automata on light field,” IEEE Transactions on Image Processing, vol. 29, pp. 1879–1889, 2019.

[11] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, “Light field saliency detection with deep convolutional networks,” IEEE Transactions on Image Processing, 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 Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 896–906.

[13] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, “Deep learning for light field saliency detection,” in IEEE International Conference on Computer Vision (ICCV), 2019, pp. 8838–8848.

[14] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency detection on light field,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2806–2813.

[15] T. Li, D. P. Lum, Y.-H. Chan et al., “Robust reflection removal based on light field imaging,” IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1798–1812, 2018.

[16] M. Le Pendu, X. Jiang, and C. Gullemot, “Light field inpainting propagation via low rank matrix completion,” IEEE Transactions on Image Processing, vol. 27, no. 6, pp. 1981–1993, 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 Winter Conference on Applications of Computer Vision (WACV). IEEE, 2020.

[18] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, “Light field image processing: An overview,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 7, pp. 926–954, 2017.

[19] S. Wanner and B. Goldluecke, “Variational light field analysis for disparity estimation and super-resolution,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 606–619, 2013.

[20] K. Mitra and A. Veeraraghavan, “Light field denoising, light field superresolution and stereo camera based refocussing using a gmms light field patch prior,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2012, pp. 22–28.

[21] R. A. Farrugia, C. Galea, and C. Gullemot, “Super resolution of light field images using linear subspace projection of patch-volumes,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 7, pp. 1058–1071, 2017.

[22] M. Rossi and P. Frossard, “Geometry-consistent light field super-resolution via graph-based regularization,” IEEE Transactions on Image Processing, vol. 27, no. 9, pp. 4207–4218, 2018.

[23] V. K. Ghassab and N. Bougila, “Light field super-resolution using edge-preserved graph-based regularization,” IEEE Transactions on Multimedia, 2019.

[24] M. Alain and A. Smolic, “Light field super-resolution via lfbm5d sparse coding,” in IEEE International Conference on Image Processing (ICIP). IEEE, 2018, pp. 2501–2505.

[25] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, “Learning a deep convolutional network for light-field image super-resolution,” in IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, 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 Processing Letters, vol. 24, no. 6, pp. 848–852, 2017.

[27] Y. Yuan, Z. Cao, and L. Su, “Light-field image super-resolution using a combined deep cnn based on epi,” IEEE Signal Processing Letters, vol. 25, no. 9, pp. 1359–1363, 2018.

[28] S. Zhang, Y. Lin, and H. Sheng, “Residual networks for light field image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 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 Transactions on Image Processing, vol. 27, no. 9, pp. 4271–4286, 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 Transactions on Image Processing, vol. 28, no. 5, pp. 2319–2330, 2018.

[31] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, “Spatial-angular interaction for light field image super-resolution,” in European Conference on Computer Vision (ECCV), 2018, pp. 144–160.

[32] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295–307, 2015.

[33] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 136–144.

[34] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in European Conference on Computer Vision (ECCV), 2018, pp. 286–301.

[35] J. Jin, J. Hou, J. Chen, and S. Kwong, “Light field spatial super-resolution via deep combinational geometry embedding and structural consistency regularization,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2260–2269.

[36] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, “A dataset and evaluation methodology for depth estimation on 4d light fields,” in Asian Conference on Computer Vision (ACCV). Springer, 2016, pp. 19–34.

[37] V. Vaish and A. Adams, “The (new) stanford light field archive,” Computer Graphics Laboratory, Stanford University, vol. 6, no. 7, 2008.
X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, “Integral human pose...”

Y. Zhao, Y. Xiong, and D. Lin, “Trajectory convolution for action...”

Y. Huang, W. Wang, and L. Wang, “Bidirectional recurrent convolutional...”

G. Bertasius, L. Torresani, and J. Shi, “Object detection in video...”

R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, “Aim 2019 challenge on constrained super-resolution: Methods and results,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 0–0.

A. Mittal, A. R. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695–4708, 2012.

A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a completely blind image quality analyzer,” IEEE Signal Processing Letters, vol. 20, no. 3, pp. 209–212, 2012.

J. Yan, J. Li, and X. Fu, “No-reference quality assessment of contrast-distorted images using contrast enhancement,” arXiv preprint, 2019.

Y. Chen, Q. Zhang, M. Lin, G. Yang, and C. He, “No-reference color image quality assessment: from entropy to perceptual quality,” EURASIP Journal on Image and Video Processing, vol. 2019, no. 1, p. 77, 2019.

M. Rerabek and T. Ebrahimi, “New light field image dataset,” in International Conference on Quality of Multimedia Experience (QoMEX), 2016.

S. Wanner, S. Meister, and B. Goldluecke, “Datasets and benchmarks for densely sampled 4d light fields,” in Vision, Modelling and Visualization (V MV), vol. 13. Citeseer, 2013, pp. 225–226.

C. Shan, H.-G. Jeon, Y. Yoon, I. So Kweon, and S. Joo Kim, “Epipet: A fully-convolutional neural network using epipolar geometry for depth from light field images,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4748–4757.

I. K. Park, K. M. Lee et al., “Robust light field depth estimation using occlusion-noise aware data costs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 10, pp. 2484–2497, 2017.

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 Transactions on Pattern Analysis and Machine Intelligence, 2019.

H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, “Depth from a light field image with learning-based matching costs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 297–310, 2018.

H. Sheng, P. Zhao, S. Zhang, J. Zhang, and D. Yang, “Occlusion-aware depth estimation for light field using multi-orientation epi,” Pattern Recognition, vol. 74, pp. 587–599, 2018.

S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, “Robust depth estimation for light field via spanning parallelgram operator,” IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3261–3273, 2019.

S. Vagharshakyan, R. Bregovic, and A. Gotchev, “Light field reconstruction using shearlet transform,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 1, pp. 133–147, 2017.

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 European Conference on Computer Vision (ECCV), 2018, pp. 137–152.

Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, “End-to-end view synthesis for light field imaging with pseudo 4dcmn,” in European Conference on Computer Vision (ECCV), 2018, pp. 333–348.

S. Zhang, H. Sheng, D. Yang, J. Zhang, and Z. Xiong, “Micro- lens-based matching for scene recovery in lenslet cameras,” IEEE Transactions on Image Processing, vol. 30, no. 1, pp. 10–18, 2021.

N. Meng, H. K.-H. So, X. Sun, and E. Lam, “High-dimensional dense residual convolutional neural network for light field reconstruction,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[86] J. Jin, J. Hou, H. Yuan, and S. Kwong, “Learning light field angular super-resolution via a geometry-aware network,” in AAAI Conference on Artificial Intelligence, 2020.

[87] N. Meng, X. Wu, J. Liu, and E. Y. Lam, “High-order residual network for light field super-resolution,” AAAI Conference on Artificial Intelligence, 2020.

[88] R. Farrugia and C. Guillemot, “Light field super-resolution using a low-rank prior and deep convolutional neural networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[89] 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 Processing: Image Communication, vol. 80, p. 115638, 2020.

[90] M. S. K. Gul and B. K. Gunturk, “Spatial and angular resolution enhancement of light fields using convolutional neural networks,” IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2146–2159, 2018.

[91] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, “Flickr1024: A large-scale dataset for stereo image super-resolution,” in IEEE International Conference on Computer Vision Workshops (ICCVW), 2019, pp. 0–0.

[92] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4681–4690.

[93] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[94] Y. Anagun, S. Isik, and E. Seke, “Srlibrary: Comparing different loss functions for super-resolution over various convolutional architectures,” Journal of Visual Communication and Image Representation, vol. 61, pp. 178–187, 2019.

[95] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, “A stereo attention module for stereo image super-resolution,” IEEE Signal Processing Letters, vol. 27, pp. 496–500, 2020.

[96] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[97] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” International Conference on Learning and Representation (ICLR), 2015.

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 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.

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 National University of Defense Technology (NUDT), Changsha, China, in 2020. He is currently pursuing the M.E. degree with the College of Electronic Science and Technology, NUDT. 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 co-authored over 100 journal and conference publications. Her current research interests include signal processing and image processing.

Yulan Guo is currently an associate professor. He received the B.E. and Ph.D. degrees from 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 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. Dr. Guo received the CAAI Outstanding Doctoral Dissertation Award in 2016, the CAAI Wu-Wenjun Outstanding AI Youth Award in 2019. He served/will serve as an associate editor for IET Computer Vision and IET Image Processing, a guest editor for IEEE TPAMI, an area chair for CVPR 2021 and IJCV 2020, an organizer for a tutorial in CVPR 2016 and a workshop in CVPR 2019. Dr. Guo is a senior member of IEEE.