Information Fusion 79 (2022) 188–199

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Learning an epipolar shift compensation for light field image super-resolution

Xinya Wang [a], Jiayi Ma [a], Peng Yi [b], Xin Tian [a,*], Junjun Jiang [c], Xiao-Ping Zhang [d]

[a] *Electronic Information School, Wuhan University, Wuhan, 430072, China*
[b] *School of Computers, Wuhan University, Wuhan, 430072, China*
[c] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China*
[d] *Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto ON M5B 2K3, Canada*

ARTICLE INFO

Keywords:
Light field
Super-resolution
Multi-view fusion
Dynamic deformable convolution

ABSTRACT

Light field imaging has drawn broad attention since the advent of practical light field capturing systems that facilitate a wide range of applications in computer vision. However, existing learning-based methods for improving the spatial resolution of light field images neglect the shifts in the sub-pixel domain that are widely used by super-resolution techniques, thus, fail in recovering rich high-frequency information. To fully exploit the shift information, our method attempts to learn an epipolar shift compensation for light field image super-resolution that allows the restored light field image to be angular coherent with the enhancement of spatial resolution. The proposed method first utilizes the rich surrounding views along some typical epipolar directions to explore the inter-view correlations. We then implement feature-level registration to capture accurate sub-pixel shifts of central view, which is constructed by the compensation module equipped with dynamic deformable convolution. Finally, the complementary information from different spatial directions is fused to provide high-frequency details for the target view. By taking each sub-aperture image as a central view, our method could be applied for light field images with any angular resolution. Extensive experiments on both synthetic and real scene datasets demonstrate the superiority of our method over the state-of-the-art qualitatively and quantitatively. Moreover, the proposed method shows good performance in preserving the inherent epipolar structures in light field images. Specifically, our LFESCN method outperforms the state-of-the-art method with about 0.7 dB (PSNR) on average.

1. Introduction

Emerging as a promising technology, light field (LF) imaging has facilitated a variety of applications, ranging from the virtual reality field to computer vision applications. Different from conventional photography, the resulting LF image records not only intensity values at each position but also directions of rays from real-world scenes. This abundant spatial–angular information makes many novel applications possible, such as post-capture refocusing [1], stereoscopic display [2] and single-shot depth sensing [3,4], especially after the advent of commercial portable LF camera (*e.g.* Lytro[1] and Raytrix[2]). As displayed in Fig. 1(a), by inserting an additional optical component like the micro-lens between the main lens and the camera sensor, the handheld plenoptic camera can capture a scene from multiple views in a single shot. The output of the camera sensor is composed of many micro-images under each micro-lens, which is enlarged in Fig. 1(b). The sub-aperture images can be further converted from the effective pixels in the camera sensor output. In Fig. 1(c), the same lines are extracted in the sub-aperture images horizontally and vertically to show the sub-pixel shifts among sub-aperture images. However, the recorded LF image is equipped with a high angular resolution at the expense of spatial resolution, which limits the range of potential development.

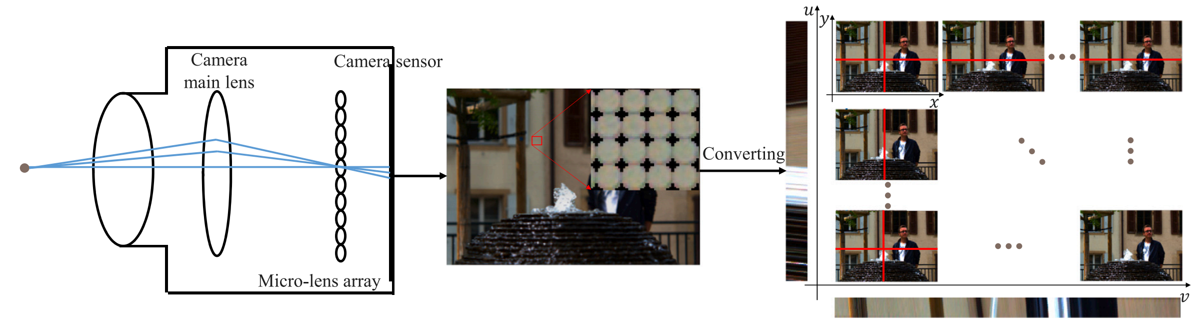
Generally, the LF image presents different view information in sub-aperture images with sub-pixel shifts in a narrow baseline so that there exist strong correlations among them, which provide the redundant data used generally by super-resolution (SR) techniques. As the internal similarity performs well in depth continuous region [5], traditional methods [6–13] for LF image SR first rely on the intrinsic imaging consistency, which explores the depth information to warp or register the sub-aperture images, and then different image priors are utilized to regularize the SR reconstruction process. Obviously, the disparity estimation is crucial for these approaches, and any defect in the depth computation or the image-level wrapping operation may introduce

* Corresponding author.
E-mail addresses: wangxinya@whu.edu.cn (X. Wang), jyma2010@gmail.com (J. Ma), yipeng@whu.edu.cn (P. Yi), xin.tian@whu.edu.cn (X. Tian), jiangjunjun@hit.edu.cn (J. Jiang), xzhang@ee.ryerson.ca (X.-P. Zhang).

[1] https://www.lytro.com/.

[2] https://www.raytrix.de/.

https://doi.org/10.1016/j.inffus.2021.10.005
Received 28 September 2020; Received in revised form 21 September 2021; Accepted 18 October 2021
Available online 1 November 2021



Object (a) Light field camera (b) Camera sensor output (c) The sub-aperture images

Fig. 1. Schematic of micro-lens based LF imaging. (a) Simple diagram of light field image. (b) Camera sensor output: we zoomed in on a small area to show the characteristics of light field imaging. (c) The sub-aperture image: to show differences between sub-aperture images (SAIs), we extracted lines horizontally (bottom of SAIs) and vertically (left of SAIs) along the red lines for visualization.

significant artifacts. Besides, the consistency is not always kept well especially in the occlusion and specular reflection regions. Recently, several learning-based methods [14–19] have been proposed for LF SR, in which the corresponding relations are implicitly excavated for different view combinations. Nevertheless, on the one hand, they are deficient in capturing the sub-pixel shifts that are provided by disparity information among different view images, resulting in the worse performance in occlusions and conterminous boundaries, as illustrated in Fig. 2. On the other hand, these methods restrict to LF images with specific angular resolution or multiple models should be trained to super-resolve single LF image.

To alleviate the above issues, we tailor a flexible and effective LF SR method in this paper, which is implemented by an epipolar shift compensation network, denoted as LFESCN. Specially, we take advantage of multiple sampling with parallax shifts in different epipolar directions, which provide redundant information generally used by super-resolution techniques. As the internal similarity performs well in the depth continuous region [5], we register multiple views from epipolar directions to the central view to capture the central sub-pixel shifts that are mapped from angular space, which contributes to the reconstruction of high-frequency details. Through explicitly exploring the coherent relations, the geometry structures that are made up by oblique lines in epipoplar plane images (EPIs) could be better preserved. To avoid the defects of explicit depth estimation and wrapping operation, the compensation module with dynamic deformable convolution is deployed for alignment in the feature domain. Concretely, taking the features from supporting and central views as input, this module could generate not only position-specific but also sample-specific offsets of deformable convolution kernels by dynamic filtering. In this way, the proposed LFESCN would have strong capability and flexibility to cope with various imaging scenes, rather than depend on imaging consistency. Finally, sub-pixel mappings from different directions are fed into the reconstruction module for feature fusion [21,22], which could provide complete residual information for the final super-resolved central view. Considering each sub-aperture image as a central view, we can easily generate the whole super-resolved LF image with any angular resolution. We conduct extensive experiments on both synthetic and real scene LF image datasets. The experimental results show that our framework achieves state-of-the-art performance.

The contributions of this paper are three-fold. (i) We propose a compensation module for feature-level registration of multiple LF views based on dynamic deformable convolution. It enables the network to explicitly capture the accurate sub-pixel shifts from epipolar directions rather than implicitly mapping. (ii) A flexible and effective framework is designed for super-resolving LF image of any angular resolution, which could preserve the inherent epipolar property better. (iii) Extensive experiments on both synthetic and real scene LF image datasets demonstrate our state-of-the-art performance compared to other LF SR methods.

2. Related work

2.1. Multi-image SR

Given multiple low-resolution images, the multi-image SR is a more general task to fuse and reconstruct one plausible image of high-resolution from unordered sets with unknown timestamps, not sequences or video. Tsai et al. [23] pioneered reconstructing a high-resolution image by fusion of low-resolution images in the Fourier domain with the assumption that their phase shifts are known. However, with the unknown shifts in practice, the fusion problem should be solved in conjunction with the registration problem [24,25].

Traditionally, optimization-based methods have assumed prior knowledge to constrain the parameter search space and derive objective functions, such as total variation [26] and Tikhonov regularization [27]. With the development of the nonparametric strategies, patch-based methods rely on sparse coding and dictionary learning to form high-resolution images directly from low-resolution patches [28,29]. Although deep learning methods are widely used in many tasks, few deep-learning approaches have attempted to solve the multi-image SR problem in an end-to-end learning framework. In [30], the first fully end-to-end architecture for multi-image SR is introduced to jointly learn and co-adapt the fusion and (co-)registration tasks to one another. Recently, several deep learning methods are proposed to tackle the multi-image SR problem without registration [31–33].

Although the light field image is composed of multiple sub-aperture images, in the light field SR, the model uses the internal information of the light field image to restore the entire high-resolution one.

2.2. LF SR

Since LF image records scene information from different angles, most LF SR methods take advantage of this rich information in the angular domain to recover missing texture details in the spatial domain. Existing methods could be divided into three categories: projection-based, optimization-based, and learning-based, referring to [5,16–18].

Projection-based method: Projection-based methods focus on projection and resample of LF data, depending on the imaging principles of light field cameras. As first introduced by Lim et al. [34], the sub-pixel shift of redundant views in angular space could be mapped into spatial space by projecting them onto convex sets, which benefits for spatial resolution enhancement. Focusing on the focal stack transformation problem, Nava et al. [35] exploited the refocusing principle and projected pixels from other views to the central view to get an all-in-focus in high resolution space. Similarly, Pérez et al. [36,37] proposed the Fourier slice super-resolution to get the super-resolved discrete focal stack transform. In [38], Georgiev and Lumsdaine established sub-pixel correspondences with the projection scheme for the focused plenotic cameras. Besides, Liang et al. [39] demonstrated that typical lenslet

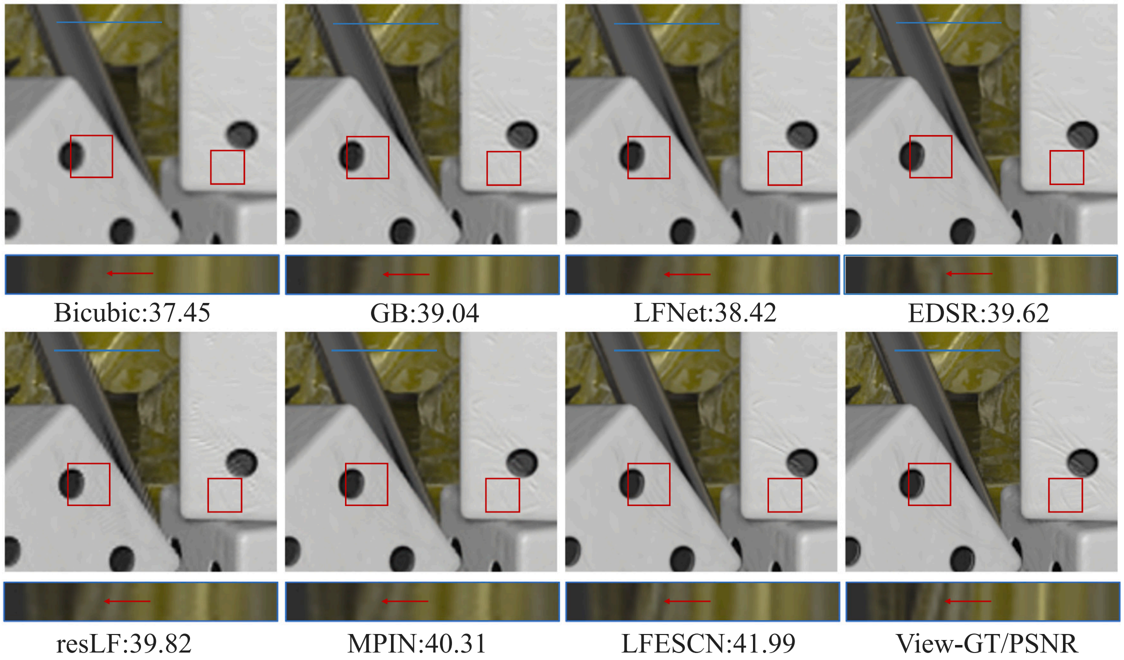


Fig. 2. The super-resolved central view of the LF image Buddha from HCI1 [20] at the upscaling factor 2. Our results of the central view images and epipolar plane images surpass the other state-of-the-art methods with higher PSNR values. As highlighted in the red box, the proposed method LFESCN could recover more texture details, especially in occlusion and reflection areas, while the others exhibit artifacts with different degrees or a faint mark. Moreover, to some extent, the inherent epipolar property in EPIs has been well preserved compared to other methods.

LF cameras preserve frequency components above the spatial Nyquist rate and performed spatial super-resolution with the guidance of depth information to project the LF samples to the target view. To relieve the dependency of camera parameters and depth information, Wang et al. [40] redefined the mapping function between the disparity of certain pixel and its shearing shift in the projection-based methods. The limitation of these methods is that only considering the internal similarity among different views could not restore rich texture.

Optimization-based method: This kind of methods first estimate depth or disparity information and rely on different priori hypotheses, thus the super-resolved LF images are found by various optimization frameworks. Bishop et al. [6,7] explicitly recovered the depth map and solved spatial light field super-resolution problem by a variational Bayesian framework with Lambertian reflectance priors. In [8], a disparity-dependent Gaussian mixture model was proposed as alternative and the super-resolved LF images were reconstructed by linear minimum mean square error estimator. Wanner and Goldluecke [41] estimated the disparity maps from the epipolar images (EPIs) with structure tensor-based method and conducted both spatial and angular SR in a variational optimization procedure. Recently, Rossi and Frossard et al. [9,10] coupled the multi-frame SR method with a graph regularizer to enforce the geometrical consistency of LF image, which avoids the explicit disparity estimation. Inspired by LFBM5D used for light field denoising, a new method proposed in [42] iteratively alternates between LFBM5D filtering and back-projection for LF SR. The performance of optimization-based methods to some extend are determined by accuracy of depth information. Furthermore, the shallow heuristic model has restricted capacity of reconstructing complex structures.

Learning-based method: Due to the extensive LF datasets, learning-based methods have emerged recently. Early, the proposed method in [43] learned a linear mapping between the LR and HR in a low dimensional subspace with ridge regression (RR). With the prosperity of deep learning, LFCNN [44] was the first CNN-based LF SR method introduced by Yoon et al. where they sent four tuples of sub-aperture image stacks into the SRCNN [45] architecture to jointly increase the spatial and angular resolution. To get better results, Fan et al. [46] developed a two-stage CNN framework, in which different sub-aperture images are aligned by patch matching in the first stage and a multi-patch fusion CNN is used in the second stage. Subsequently, a shallow CNN model was used to super-resolve the LF raw data directly from plenoptic cameras without decoding to sub-aperture images in [47]. Regarding an LF image as a sequence of 2D images, LFNet [14] was developed to model the spatial correlation between adjacent views in a bidirectional recurrent way and accumulated contextual information from multiple scales with a specially designed fusion layer. With a combined CNN architecture, Yuan et al. [15] designed an EPI enhancement network as the post-procedure of EDSR [48] conducting on single view of LF image. Inspired by epipolar geometry used for depth estimation, Zhang et al. [16] grouped different image stacks into different branches to super-resolve the central view by residual learning in several position-specific model. In [17], each view of an LF image is first individually super-resolved by exploring the complementary information and the whole parallax relationship is enforced by a regularization network. Recently, Wang et al. [18] proposed a spatial–angular interactive network to extract spatial and angular features separately. In [19], the 4D LF image is rearranged into the 2D macro-pixel image to fully exploit spatial–angular correlations. However, most of them neglect sub-pixel shifts in multiple angular directions or implicitly exploit the relations by stacking them together. Therefore, our proposed method makes full use of the view information from all directions by deploying the dynamic deformable convolution in the compensation module to capture the sub-pixel shifts explicitly. In this way, our method can not only recover rich details even in handling complex scenes but also performs well in preserving the epipolar property.

2.3. Deformable convolution

Equipped with a regular grid of the filter configuration, the CNN has the inherent limitation in modeling irregular geometric transformations. For improving the transformation modeling capability of regular CNN, Dai et al. [49] proposed a deformable convolution operation that augments the spatial sampling locations in convolution with additional offsets and learns the offsets from target tasks. In practice, an extra convolution is applied to model the offset function, allowing the deformation to condition on the input in a local, dense and learnable manner. Hence, it can adaptively learn to model various transformations. In addition, the deformable convolution is further improved in [50] which introduces the modulation mechanism into the standard deformable module to strengthen the capability in manipulating spatial support regions. Recently, the deformable convolution has been proven to be superior on many high-level vision tasks such as object detection [51,52], semantic segmentation [53] and human pose estimation [54]. Besides, the deformable convolution is also succeeded in video SR problem [55–58]. Inspired by [55] which used it to implement temporal alignment, we design the compensation module to map the sub-pixel shifts from angular space to spatial space, which is demonstrated to be beneficial for the enhancement of spatial resolution.

3. Methodology

3.1. Overview

Since an LF image records a scene in multiple views with different angular resolutions, it is unrealistic to take all sub-aperture images into consideration when only working on improving spatial resolution. The existing learning-based algorithms for LF SR are dedicated to constructing complicated frameworks to super-resolve the whole LF image through either generating sub-aperture images recurrently in the vertical and horizontal ways [14] or adopting different models to the view in the specific location [16]. For sub-aperture images, they cannot be treated equally because of the unbalanced auxiliary information or the specific order. In this way, the performance of edge views would suffer from decay to some extent. Meanwhile, for LF images with different angular resolutions, these methods should be trained from scratch. To this end, as shown in Fig. 3, we consider the multiple views from different directions as the epipolar constraint to super-resolve the central view. Ideally, in the case of Lambertian hypothesis, the internal correlation among different sub-aperture images appears as imaging consistency when recording the same scene. That is, the surrounding view can be obtained from the central view with disparity information and its viewing location. Owing to the regular arrangement of the micro-lens, the disparity information among the LF image can be approximately modeled in a linear relationship [5]. Consequently, we stack multiple views from the same directions as auxiliary inputs. In order to fully exploit inter-view correlations, the compensation module is designed for feature-level registration of multiple LF views, which is supposed to map the parallax shifts in all epipolar directions from angular space to spatial space. In this way, our method could capture more accurate sub-pixel shifts that are beneficial for the high-frequency information recovery. By taking each view as the central view, our method could handle the LF image with different angular resolutions flexibly.

Suppose $L^{LR} \in \mathbb{R}^{U \times V \times H \times W \times C}$ is the low-resolution (LR) LF image, and $L^{HR} \in \mathbb{R}^{U \times V \times sH \times sW \times C}$ is the corresponding high-resolution (HR) LF image with the upscale factor s. The goal of our method is to reconstruct each sub-aperture image $I_{u,v}^{SR}$ from the reference LR viewpoint $I_{u,v}^{LR}$ and supporting views along epipolar directions. Generally, the strictly equidistant lattice configuration of the sub aperture image results in the parallax of light field image along each angular dimension. Thereby, as illustrated in Fig. 3, we divide multiple views into four different stacks $\{I_{0^\circ}^{LR}, I_{45^\circ}^{LR}, I_{90^\circ}^{LR}, I_{135^\circ}^{LR}\}$ according to relative angular position and each stake concatenates six sub-aperture images directly because of the narrow baseline. Thus, our LFESCN framework adopts four LR stacks as auxiliary information to predict the central HR view:

$$I_c^{SR} = \mathcal{F}_{ESCN}(I_c^{LR}, I_{0^\circ}^{LR}, I_{45^\circ}^{LR}, I_{90^\circ}^{LR}, I_{135^\circ}^{LR}; \theta), \tag{1}$$

where θ represents the parameters of our method. Through changing (u, v) coordinate of the central view, we super-resolve each sub-aperture image separately to generate the whole LF image with the enhancement of spatial information. For the edge sub-aperture image with some missing supporting views, we would reuse the existing views according to the corresponding angular direction to form a complete input stack, as shown in the bottom left of Fig. 3. In this way, all LR views are supported by equivalent view information, recovering as much texture as possible.

3.2. Network design

Along one angular direction, the multiple viewpoints embody sub-pixel shifts from a specific direction in the spatial domain because of the occlusion and angle of view. Based on this observation, we construct multi-stream branches to process each image stack separately at the beginning of our network. Subsequently, in order to make full use of both intra-view spatial correlations and inter-view angular correlations, each branch is aligned with the central target view in the feature domain. Due to the different occlusions and photometric changes in the scene, a specific filter is generated for each spatial position of the input to cope with different deformations. Instead of optical estimation for motion, the offsets of sampling convolution kernels are more suitable for the changes in a narrow baseline. In this manner, the disparity information is implicitly exploited to find out accurate sub-pixel shifts between each stack and the target view, which would be mapped into residuals from different directions.

As depicted in Fig. 3, our LFESCN network takes five branches as input. For the target view, after one convolution layer to extract the shallow feature, we adopt a feature extraction module to acquire high-level information, which can be formulated as:

$$F_c = f_c(I_c^{LR}), \tag{2}$$

$$M_c = H_e(F_c), \tag{3}$$

in which f_c denotes the common convolution layer operated on the central view and H_e is operation of the extraction module. For other branches, the corresponding features are generated by the same way:

$$M_a = H_e(f_a(I_a^{LR})), a \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}, \tag{4}$$

where f_a denotes the common convolution layer operated on the supporting branches and the subscript a represents a specific angular direction, that is $a \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. Since the simplified residual block [48] shows the outstanding performance in SR task, we employ k blocks with m feature maps in the extraction module to extract the rich feature representation. The extracted features will be utilized for feature-wise registration.

Afterwards, we apply a compensation module to features from one auxiliary branch and the central branch, which is supposed to perform registration in the feature domain to get accurate spatial sub-pixel shifts. Thus, we would have the aligned features M'_a from one angular direction to the target view as follows:

$$M'_a = H_c(M_a, M_c), a \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}, \tag{5}$$

in which H_c represents the compensation module, as detailed in Section 3.3.

After achieving the alignment of sub-pixel shifts, we fuse all the features in a concatenating way and further feed them into the reconstruction module. Since the different branch contains the shift from a specific direction, we integrate the fusing shift information in the reconstruction module to produce the residual of central view:

$$R_c = H_r([M_c, M'_{0^\circ}, M'_{45^\circ}, M'_{90^\circ}, M'_{135^\circ}]), \tag{6}$$

where H_r denotes the reconstruction module. In this module, the stacked aligned features are taken as input to predict the deep features,

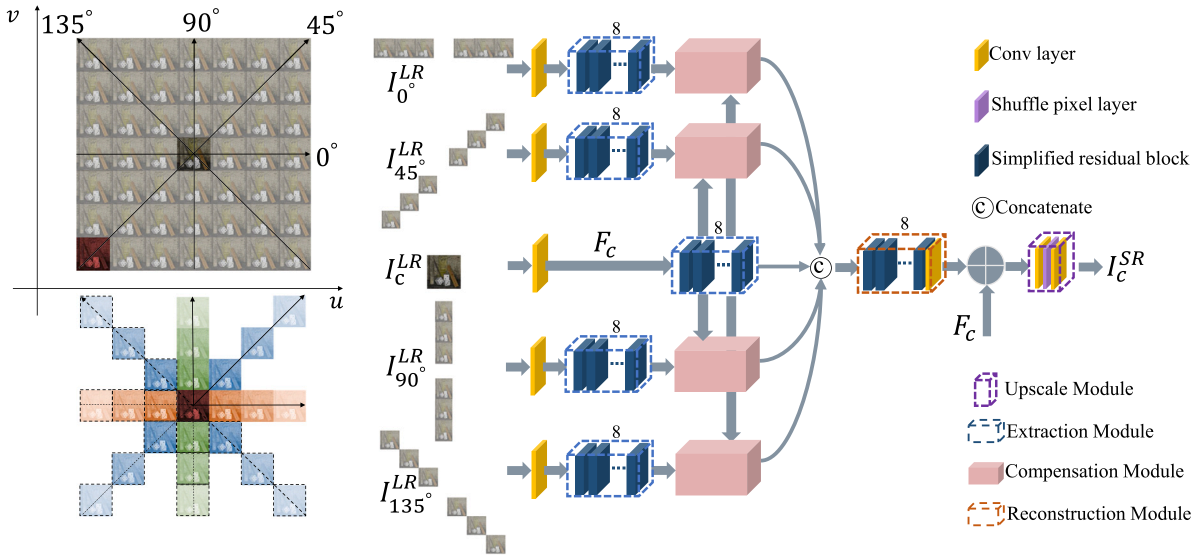


Fig. 3. The overall structure of the proposed LFESCN network. For the central view I_c^{LR}, along different directions, six images are stacked into I_a^{LR}, $a \in \{0°, 45°, 90°, 135°\}$ as epipolar constraint to super-resolve the central view. By taking each view as the central view, our method could reconstruct the high-resolution LF image flexibly. For the edge sub-aperture image with some missing surrounding views, existing views are reused according to the corresponding angular direction to form a complete input stack. The bottom left plot shows an example of this procedure for the bottom left sub-aperture image with a red mask. The views with masks of the same color are the same sub-aperture image, and the one with a dotted box is a duplicate of the corresponding original sub-aperture image.

and k residual blocks followed by one convolutional layer are used for fusing the feature shifts from different directions. Therefore, the global residual would possess auxiliary information from all directions to yield the HR central view. Therefore, we can produce the super-resolved central view by:

$$I_c^{SR} = H_{up}(F_c + R_c), \tag{7}$$

among which H_{up} represents the upscaling module to improve the spatial resolution. Same as the common upscaling module in single image SR network, we adopt an upscaling layer to increase the resolution of the feature map with a sub-pixel convolution as proposed by Shi et al. [59].

In this way, our method would fully excavate the angular correlations among sub-aperture images to aid the super-resolution process and meanwhile preserve the inherent properties of LF image.

3.3. Compensation module

In contrast to the traditional method of depth estimation followed by wrapping, we realize the alignment on feature level to implicitly utilize parallax information. In the video SR task, optical flow is widely estimated to capture the motion from the reference frame to the target frame. However, in the LF image, there is no large movement between sub-aperture images, but parallax shifts of several pixels due to occlusion and angle of view. Accordingly, we tailor a compensation module to capture the accurate sub-pixel shifts by registering the supporting views to the central view, where the position-specific filters are learned to produce the offsets of specific supporting views for the deformable convolution, as shown in Fig. 4.

In the compensation module, taking the M_a and M_c as input, we aim to predict the corresponding aligned LR feature M'_a for the target view. Inspired by [60], first, we generate the filters for every position of features from the adjacent views via a filter-generating network as follows:

$$W_a = H_{fg}([M_c, M_a]), \tag{8}$$

where W_a is the weight of the learned filter applied for the supporting views and H_{fg} represents the filter-generating network that shares the same insight from U-Net. Particularly, the spatial size of W_a is the same as the input so that the filtering operation is not translation invariant anymore. Instead, different filters are applied to the specific positions of M_a to acquire the offset. For each position (i,j) of the input M_a, a specific local filter $W_a^{(i,j)}$ is applied to the region centered around $M_a^{(i,j)}$, which is expected to produce the offset:

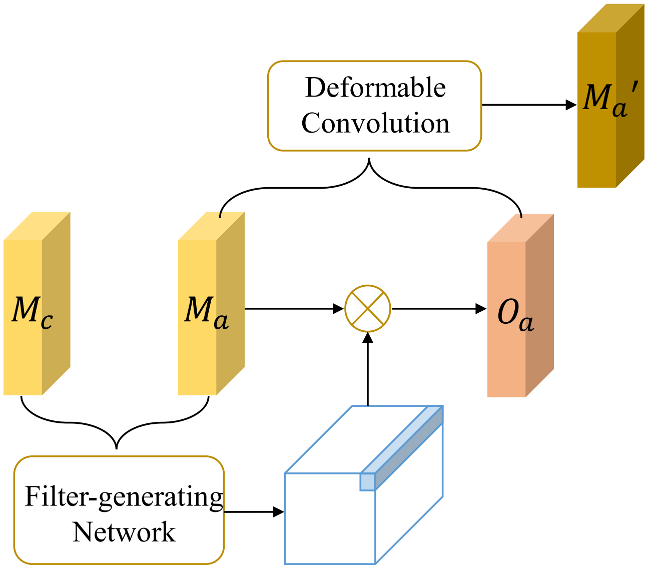


Fig. 4. The compensation module of our proposed LFESCN.

$$O_a^{(i,j)} = \sum W_a^{(i,j)} M_a^{(i,j)}. \tag{9}$$

In this way, the filters used in this layer are not only sample specific but also position specific to deal with the photometric change. Subsequently, the deformable layer could be applied for the corresponding features and we can obtain the aligned features from:

$$M'_a = f_{dc}(M_a, O_a), \tag{10}$$

in which f_{dc} denotes the deformable convolution. As the common convolutional layer has a regular grid of a 3 × 3 kernel size, *e.g.*

$\mathcal{R} = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$, O_a would offer the offsets of deformable convolution kernels, that is $O_a = \{\triangle\ p_n | n = 1,\ldots,|\mathcal{R}|\}$. Therefore, for each location p_0 on the aligned feature maps M'_a, we have:

$$M'_a(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) M_a(p_0 + p_n + \triangle\ p_n), \tag{11}$$

where p_n enumerates the locations in $\mathcal{R}$. In this process, the sampling kernel is operated on irregular grid $p_n + \triangle\ p_n$, where $\triangle\ p_n$ might be fractional. Same as that proposed in [49], Eq. (11) is implemented by bilinear interpolation. Through an irregular respective field, this compensation module could capture the sub-pixel shift from all epipolar directions, which would benefit for the SR process.

3.4. Implementation details

For both the extraction and reconstruction modules, each residual block has two convolutional layers with a rectified linear unit inserting between them. In all convolutional layers, we choose 3×3 as the kernel size and pad zero to avoid border effects. Experimentally, we set $k = 8$ and $m = 64$. Thus, in the reconstruction module, the layers have 64×5 feature maps after concatenating features from all directions. To control model size, five branches share the parameter except for the first convolutional layer. In the filter-generating network, six layers are deployed with the filter numbers of $\{64, 32, 16, 32, 64, 64 \times 9\}$. Since the last layer produces the position-specific filters for the features from supporting views, each position is equipped with the 3×3 respective field. Our network deals with RGB images so that the input and output channel is 3 and we use the $L1$ loss function to generate better performance compared to $L2$ loss.

For the training phase, we empirically set a mini-batch size of 16 with the spatial size of 48×48 as inputs and employ Adam optimizer with weight decay of $1e{-}4$ to train our model. All weights of the layers in our network are initialized by Xaviers algorithm [61]. The learning rate is initialized as $1e{-}3$ and is decayed by 10 times after 100 epochs until the validation loss converges. Our model is implemented by Pytorch on NVIDIA GTX 1080Ti. An LF image with the resolution of $512 \times 512 \times 7 \times 7 \times 3$ can be spatially super-resolved within 1.5 s at the scale 2, roughly 0.03s per sub-aperture image. The source code and experimental datasets to reproduce the super-resolved results will be released upon the acceptance of submission.

4. Experiments and results

4.1. Datasets and settings

To validate the effectiveness of our proposed LF SR method, we conduct extensive experiments on both synthetic and real scene light field datasets. The synthetic LF images from [20,62] and real scene LF images [63–65] from Lytro Illum cameras are collected as a dataset, which contains rich LF scenes, various in spatial and angular resolution. Specifically, for a fair comparison, we select Buddha and Mona in the synthetic LF dataset HCI1 [20] into the test set on purpose. The remaining images are randomly divided: two images for validation and six images for training. In the synthetic LF dataset HCI2 [62], we use pre-divided LF images and there are 16, 4, and 4 LF images in training, validation, and test sets, respectively. For the real scene EPFL [63] dataset, we choose specific 12 images for testing as most existing methods do [14,16] and randomly divide 12 images for validation and the resting 84 images for training. There are 9 kinds of imaging scenes in the real scene Stanford dataset [64]. For each scene, we select 14 high-quality images and randomly divide 10, 2, and 2 images into training, validation, and test sets. The real scene DDFF dataset [65] has 6 kinds of imaging scenes and each scene has 10 images, which are randomly divided into 6, 2, and 2 samples for training, validation, and validation for each scene. Consequently, there are in total 250, 48, 48 images in training, validation and test datasets without overlap. For any dataset, LF images are cropped with 7×7 angular resolution free from the border effects and then regarded as ground-truth images. Specifically, we downsample them spatially at the scaling factor 2, 3 and 4 by bicubic interpolation to acquire the LR input. The super-resolved images are evaluated by widely used measurements: PSNR and SSIM.

Table 1
The performance and speed of different fusion operators on the test dataset for the scale factor 2.

Fusion operator	Addition	Multiplication	Pooling	Concatenation (Ours)
PSNR (dB)	40.38	40.24	40.32	40.68
Time (s)	0.98	1.02	1.13	1.28

4.2. Ablation study

In order to demonstrate the effectiveness of our proposed LFESCN, we compare it with the baseline: EDSR [48], which represents the single image SR model with only target view as input. We train the EDSR model on the same LF dataset from scratch. In the existing works [49,55] that involve deformable convolution, more deformable layers are proved to enhance the capability of the network. Therefore, we further investigate our LFESCN models with different numbers: 0, 1, and 2 of compensation modules, which denote as LFESCNCMX. Without the compensation module, the network LFESCNCM0 just adopts the adjacent views in five branches to super-resolve the central view, similar as the multi-frame SR task.

Effectiveness of the compensation module. Fig. 5 illustrates convergence curves of the aforementioned models, evaluated on the validation set. We could observe: the performance of LFESCNCM0 is better than that of EDSR, which indicates that exploiting adjacent views even without shift compensation could improve the LF SR performance. Besides, the proposed method with compensation modules obtains further improvements. The proposed shift compensation network is more effective in utilizing the information from supporting views. However, opposed to other tasks involving deformable convolution, more compensation modules do not enhance the capacity of our LFESCN. Considering the narrow baseline in sub-aperture images of LF, the parallax is generally around a few pixels according to the scene. Thus, more compensation modules that could capture long-distance dependency or large motion is redundant in the LF SR task.

Effectiveness of the fusion operator. In our proposed method, we fuse all the features in a concatenating way and further feed them into the reconstruction module. To verify the effectiveness of this fusion operator, we replace the concatenation operation with addition, multiplication, and pooling. The performance and speed of these variants are evaluated on the test dataset at the scale factor 2, which is demonstrated in Table 1. For the addition and the multiplication operations, we directly replace the concatenation in Fig. 3. Although the use of addition and multiplication can reduce the burden of the network to a certain extent, these two fusion patterns cannot make full use of the extracted auxiliary information, resulting in poor reconstruction performance. For pooling operation, we deploy 1×1 convolution layer for channel reduction and fusion to replace the concatenation operation and the “pooling" method could achieve better results due to the trainable advantage. Compared with other ways of fusion, the concatenation operator acquires a large performance gain with a slower running time. However, our method is proved to be faster than other compared methods.

Visualization of sampling positions. To further verify the effectiveness of the compensation module, we visualize the sampling positions based on learned offsets to demonstrate how the compensation module explicitly capture the accurate sub-pixel shifts. As displayed in Fig. 6, two representative regions (3×3 pixels) from each central view

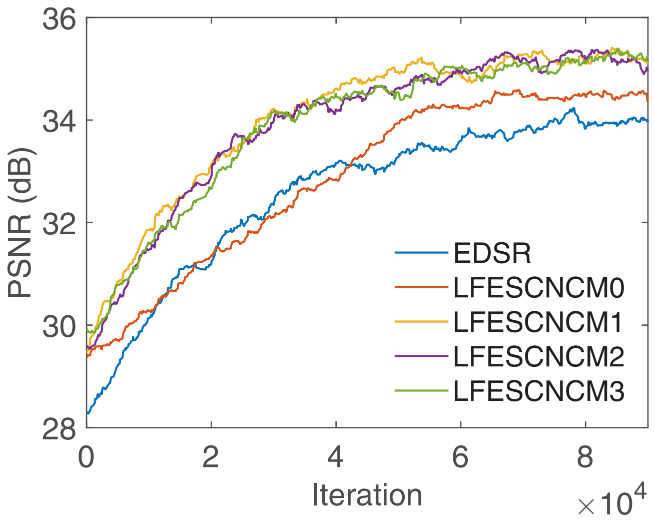


Fig. 5. Convergence analysis of LFESCN with different numbers of compensation modules and one baseline: EDSR [48]. The curves show PSNR values of different models on the validation dataset for 9×10^4 iterations.

of the LR LF image are selected and we show sampling positions on the last views from each supporting stack rather than feature maps for better visualization. Since one layer with 3 × 3 kernels is used to sample features from surrounding views for feature-level registration, each output pixel corresponds to 9 learned sampling points. We can see that sampling positions tend to capture visual regions of different shapes containing similar content rather than scanning over the whole objects. Besides, for views from different stacks, the compensation module can adaptively capture the shifts in different directions.

Influence of sub-aperture images. We also investigate the influence of sub-aperture images with different numbers. For a fair comparison, the number of parameters and the architectures of the networks are kept the same with the original network except for the input channel of the first convolution layer. The quantitative comparisons on the test dataset at the scale factor 2 are illustrated in Table 2. In can be seen that even only with two views in a stack, our results are better than that of resLF [16] (39.64/0.9842). Besides, both more auxiliary stacks and more views from the same direction could bring greater performance improvements. Nevertheless, with the increase in distance, the performance improvement of more auxiliary inputs decreases gradually. As most LF images have 7 × 7 angular resolution, we use six views in a stack to get the best performance.

4.3. Comparison methods

For comprehensive comparison, the results of representative methods are compared, including three categories: (1) single image SR methods: bicubic interpolation (Bicubic) and EDSR [48]; (2) optimization-based LF SR method: GB [9]; (3) learning-based LF SR methods: RR [43], LFNet [14], resLF [16], and MPIN [19]. As the EDSR method is originally trained for single image SR, we train the network on the same LF dataset from scratch, following the protocol in [48]. GB [9] method is set the same parameters as in the original paper. For RR [43], we use the PCA basis and the learned transformation matrix provided by the authors. As the code of LFNet is released based on Theano, we use it with the same parameters provided in the paper. Besides, we use the released models in resLF for testing, and the MPIN method only performs at scale 2 and 4 according to the original paper.

Due to the fact that the input channel varies in different methods, we calculate the quantitative index on the Y channel of the SR image in the YCbCr space.

4.4. Synthetic dataset

As the test dataset consists of a variety of LF scenes, we divide them into two categories: synthetic dataset and real scene dataset.

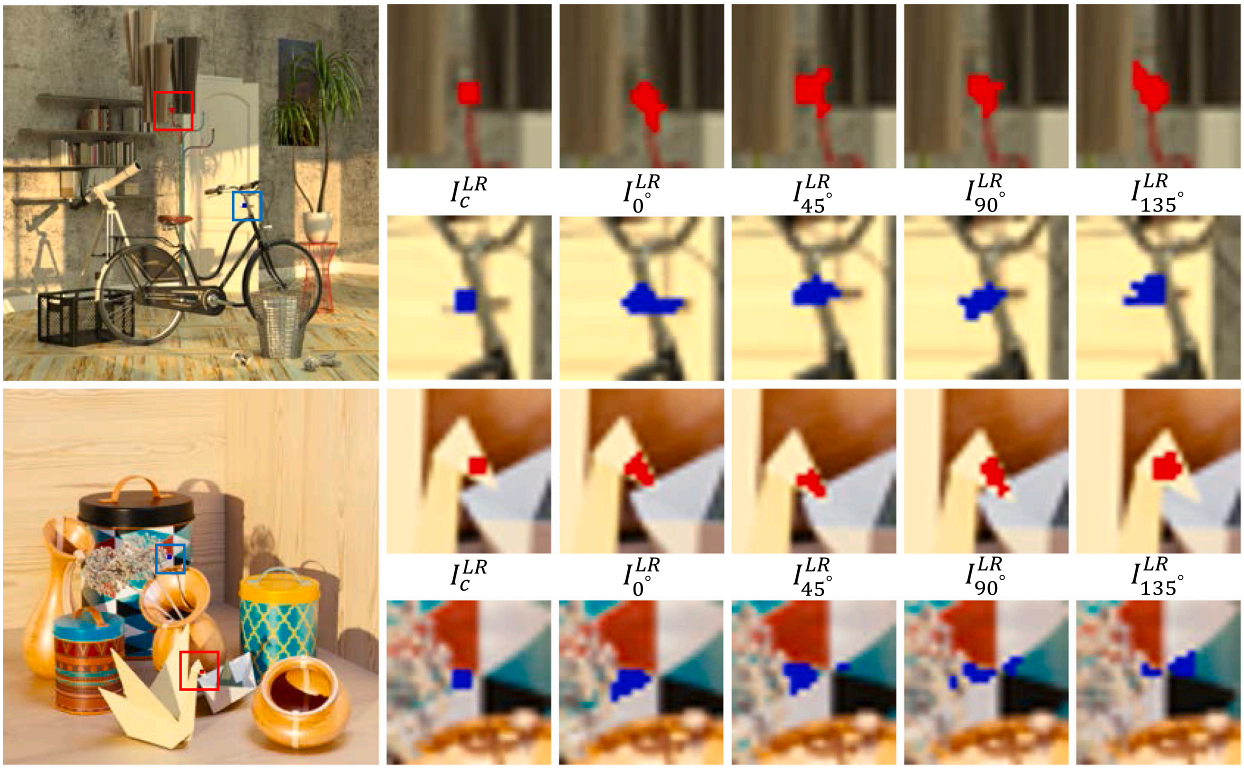


Fig. 6. Visualization of the learned sampling positions. For the 3 × 3 regions in the central view, we display the sampling positions on the last views from different directions.

Table 2
Quantitative comparisons using different numbers of sub-aperture images on test dataset for the scale factor 2. Bold indicates the best result.

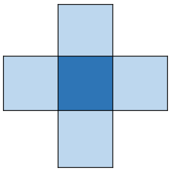
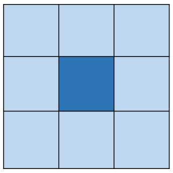
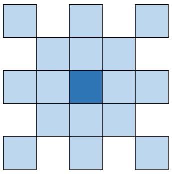
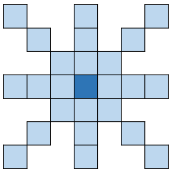
Input				
Avg.PSNR	39.27	40.08	40.56	**40.68**
Avg.SSIM	0.9839	0.9861	0.9866	**0.9867**

Table 3
Detailed comparison on HCI1 dataset of state-of-the-art LF SR algorithms: minimum, average and maximum PSNR/SSIM for scale factors 2. Bold indicates the best result.

Method	Buddha			Mona		
	Min	Avg	Max	Min	Avg	Max
Bicubic	37.24/0.9445	37.45/0.9464	37.76/0.9488	37.28/0.9520	37.42/0.9529	37.53/0.9536
RR [43]	37.81/0.9497	37.99/0.9518	38.31/0.9546	38.09/0.9573	38.27/0.9592	38.46/0.9626
GB [9]	38.57/0.9582	39.04/0.9606	39.29/0.9626	38.81/0.9649	39.05/0.9663	39.22/0.9675
LFNet [14]	38.09/0.9709	38.42/0.9731	38.77/0.9760	38.38/0.9884	38.73/0.9891	38.80/0.9895
resLF [16]	38.71/0.9744	39.82/0.9825	40.94/0.9866	39.30/0.9809	41.19/0.9879	42.20/0.9907
EDSR [48]	39.31/0.9827	39.62/0.9837	39.99/0.9846	41.15/0.9890	41.32/0.9892	41.47/0.9894
MPIN [19]	39.78/0.9842	40.31/0.9864	40.70/0.9878	41.93/0.9904	42.25/0.9912	42.53/0.9919
LFESCN	**41.46/0.9901**	**41.99/0.9907**	**42.28/0.9913**	**43.32/0.9932**	**43.46/0.9935**	**43.64/0.9937**

Table 4
Quantitative evaluation on test dataset of state-of-the-art LF SR algorithms: average PSNR/SSIM for scale factors 2, 3 and 4. Bold indicates the best result.

Method	Scale	Synthetic dataset		Real scene dataset		
		HCI1 [20]	HCI2 [62]	EPFL [63]	Stanford [64]	DDFF [65]
Bicubic	×2	37.43/0.9497	32.89/0.8903	32.42/0.9275	37.24/0.9539	35.45/0.9240
RR [43]	×2	38.13/0.9555	33.40/0.8979	33.54/0.9339	37.46/0.9547	35.96/0.9230
GB [9]	×2	39.04/0.9634	34.96/0.9278	31.45/0.8896	38.82/0.9647	36.15/0.9293
LFNet [14]	×2	38.57/0.9811	33.73/0.9544	33.73/0.9715	38.26/0.9834	36.72/0.9718
resLF [16]	×2	40.50/0.9852	36.38/0.9764	37.33/0.9815	42.53/0.9909	38.58/0.9780
EDSR [48]	×2	40.47/0.9864	35.48/0.9656	37.20/0.9788	41.47/0.9897	38.69/0.9790
MPIN [19]	×2	41.28/0.9888	36.89/**0.9776**	37.68/0.9816	41.93/0.9923	38.37/0.9768
LFESCN	×2	**42.73/0.9920**	**37.34/0.9776**	**39.22/0.9862**	**43.17/0.9926**	**39.16/0.9804**
Bicubic	×3	34.31/0.8980	30.27/0.8148	29.74/0.8677	33.79/0.9046	33.31/0.8635
RR [43]	×3	35.18/0.9136	30.91/0.8329	30.77/0.8845	34.34/0.9127	32.81/0.8711
GB [9]	×3	35.43/0.9187	31.55/0.8535	30.59/0.8896	34.81/0.9211	32.74/0.8724
LFNet [14]	×3	34.88/0.9597	30.58/0.9124	30.28/0.9391	34.23/0.9575	32.96/0.9397
EDSR [48]	×3	36.80/0.9683	32.09/0.9309	33.98/0.9612	36.78/0.9716	34.73/0.9531
LFESCN	×3	**38.26/0.9778**	**33.37/0.9482**	**34.62/0.9690**	**38.14/0.9782**	**35.12/0.9564**
Bicubic	×4	32.40/0.8499	28.82/0.7599	28.15/0.8162	31.86/0.8616	30.44/0.8130
RR [43]	×4	33.24/0.8702	29.40/0.7788	29.00/0.8354	32.36/0.8708	30.91/0.8221
GB [9]	×4	33.37/0.8741	29.75/0.7939	28.84/0.8380	32.62/0.8783	30.89/0.8238
LFNet [14]	×4	33.14/0.9396	29.30/0.8836	28.90/0.9125	32.49/0.9365	31.15/0.9107
resLF [16]	×4	34.93/0.9506	30.65/0.9134	31.27/0.9405	34.30/0.9530	32.15/0.9268
EDSR [48]	×4	34.55/0.9471	30.33/0.9002	32.03/0.9396	34.15/0.9510	32.38/0.9245
MPIN [19]	×4	35.56/0.9604	31.12/0.9186	31.41/0.9456	34.65/0.9552	32.58/0.9294
LFESCN	×4	**35.89/0.9623**	**31.14/0.9189**	**32.45/0.9472**	**35.02/0.9586**	**32.77/0.9302**

HCI1 [20] and HCI2 [62] are the synthetic datasets, which contain 2 and 4 samples, respectively. Specially, Table 3 exhibits the detailed comparisons of image Buddha and Mona in HCI1 [20], where the minimum, average and maximum PSNR and SSIM at the magnification factor 2 are provided. It can be noticed that the result of resLF varies greatly in different views, where the maximum exceeds the minimum nearly 3 dB of PSNR value even if the average value is higher than that of EDSR. Since in resLF, the model used for SR differs in view position and sample number, the auxiliary information is unbalanced for each view, and thus resLF suffers from diversity in performance among different views. Our method outperforms the other methods with more than 1.5 dB (PSNR) in Buddha and 1.2 dB (PSNR) in Mona. Although the differences in our super-resolved view images are bigger than that of EDSR, the minimum values of our method are still higher than their maximum values. In the meanwhile, the first two columns of Table 4 illustrate the average quantitative indexes of six methods evaluated on the synthetic dataset at the upscaling factors 2, 3 and 4. Based on the results, our proposed method still has advantages over the state-of-the-art method. We could observe that the deep learning-based method LFNet performs even worse than the optimized-based method GB. The reason might be that LFNet learns the super-resolved images from horizontal and vertical stacks, resulting in insufficient integration from different directions. Whereas, in the proposed LFESCN, we utilize multiple views to provide the sub-pixel shift for the central view. Moreover, the internal similarity is exploited explicitly along all epipolar directions to integrate sub-pixel shifts, which contributes to the improvement of spatial resolution.

To demonstrate the effectiveness of our proposed SR method for synthetic LF data, we also display the reconstruction results of compared methods in Fig. 7 with the ground truth in the last column. Two representative images are selected: Mona from HCI1 [20] and bedroom from HCI2 [62] at the scale factor 4 that is regarded as harder SR task.

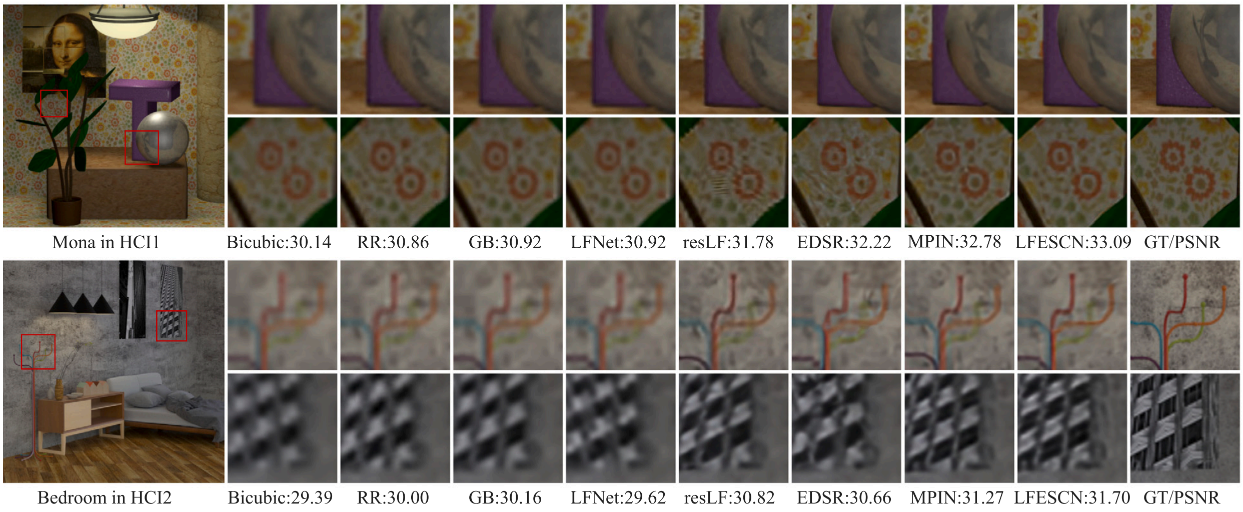


Fig. 7. The detailed ×4 super-resolution results for synthetic images Mona and Bedroom. The super-resolved central view images are shown, where the corresponding PSNR values of central views are illustrated below.

As shown in Fig. 7, our method could recover more texture details or clear structures which are blurry or destroyed by severe artifacts in other reconstruction results. Especially, we could observe that even in reflection surface, our method also behaves well in acquiring gradations of light and shadow. Owing to the dynamic deformable convolution in the compensation module, the proposed method could capture the photometric change from rich views.

4.5. Real scene dataset

The LF images of real scene dataset are captured using plenoptic cameras. According to their sources, the real scene dataset is classified into three subsets: EPFL from [63], Stanford from [64] and DDFF from [65], where there are 12, 18 and 12 LF images, respectively. Evaluated on real scene datasets, the last three columns of Table 4 compare the average quantitative results of different methods at the magnification factors 2, 3 and 4. As described in Table 4, the performance of resLF suppresses that of EDSR on EPFL and Stanford when upsampling LF images two times. This is mainly because this two subsets have too many specific LF imaging scenes, such as reflective surfaces and occlusions, for which multi-view information would helpful. It can be seen that owing to the strong capability of deep network, EDSR achieves competitive performance despite that it is not designed for LF image. With the guidance of adjacent views, our proposed LFESCN outperforms the others in terms of two qualitative indexes by learning the epipolar shift compensation. Since the view images of real world scenes contain the interference of noticeable artifacts and noises, the original LF images in this dataset are relatively of low quality. Under the circumstances, our method still shows its superiority. On the contrary, attributed to sub-pixel information learned from multiple views, it demonstrates that our method could handle challenging scenarios much better than state-of-the-art.

For visual comparison of real scene dataset, Fig. 8 shows the central view of SR image reconstructed by compared methods at the upscale factor 4. We choose three LF images from the subsets, respectively, including Bikes from EFPL [63], Fruits and vegetables from Stanford [64] and Library from DDFF [65]. In Fig. 8, we enlarge the area in the red box with the ground truth at the left-down of whole image, where the PSNR value evaluated on the central view is shown below. It can be seen that the results of LFNet are corrupted by obvious artifacts, although this method is especially trained on real scene images. EDSR produces ambiguous even over-smoothed super-resolved results as the redundant view information is not explored. By contrast, our results are relatively clear and real in different kinds of real scene images, which demonstrates that the proposed method could not only recover more high-frequency details but also deal with complex imaging scenes.

Table 5
Quantitative evaluation of EPIs on test dataset reconstructed by LF SR algorithms: average SSIM for scale factor 4. Bold indicates the best result.

Method	Synthetic dataset	Real scene dataset
bicubic	0.8269	0.8602
RR [43]	0.8366	0.8630
GB [9]	0.8485	0.8705
LFNet [14]	0.8425	0.8685
resLF [16]	0.8547	0.8719
EDSR [48]	0.8501	0.8911
MPIN [19]	0.8592	0.8826
LFESCN	**0.8681**	**0.8936**

4.6. EPIs comparison

Since we super-resolve each sub-aperture image separately, it is vital to verify whether our method could preserve the inherent geometric structure. For the sake of evaluating the epipolar property, we transform the predictive LF images into EPIs horizontally and vertically and calculate the average SSIM of different methods, which is displayed in Table 5. According to the results, our method achieves the best performance on both synthetic and real scene datasets. Thus, to some extent, the proposed method could preserve much more epipolar geometric structures in terms of EPIs.

To further show our advantages in preserving the inherent epipolar property, we also visualize the EPI results in Figs. 2 and 8 along the line in the picture. As depicted in the visualization results, the reconstructed EPIs of RR and GB are blurry. Although the image stacks are simultaneously super-resolved in LFNet, the oblique lines are still distorted since the epipolar constraint from all directions is not fully exploited in the super-resolution process. Due to the fact that EDSR is not designed for LF SR tasks, it suffers from distortion in EPIs despite the high PSNR value on a single image. Compared to other methods, by considering the surrounding view images as epipolar constraint, our LFESCN could recover more sharp lines in EPIs. In the meanwhile, the dynamic deformable layer in the compensation module makes our method sensitive to the light change so that the inherent geometric structure is well preserved in our reconstructed LF images. The basic idea of the proposed method is to map the sub-pixel shift from the angular dimension to the spatial dimension along all epipolar directions through alignment in the feature domain. Therefore, on the one hand, using all surrounding views as priori information, it can be ensured that the inherent epipolar property is largely preserved during the super-resolution process. On the other hand, owing to the explicit

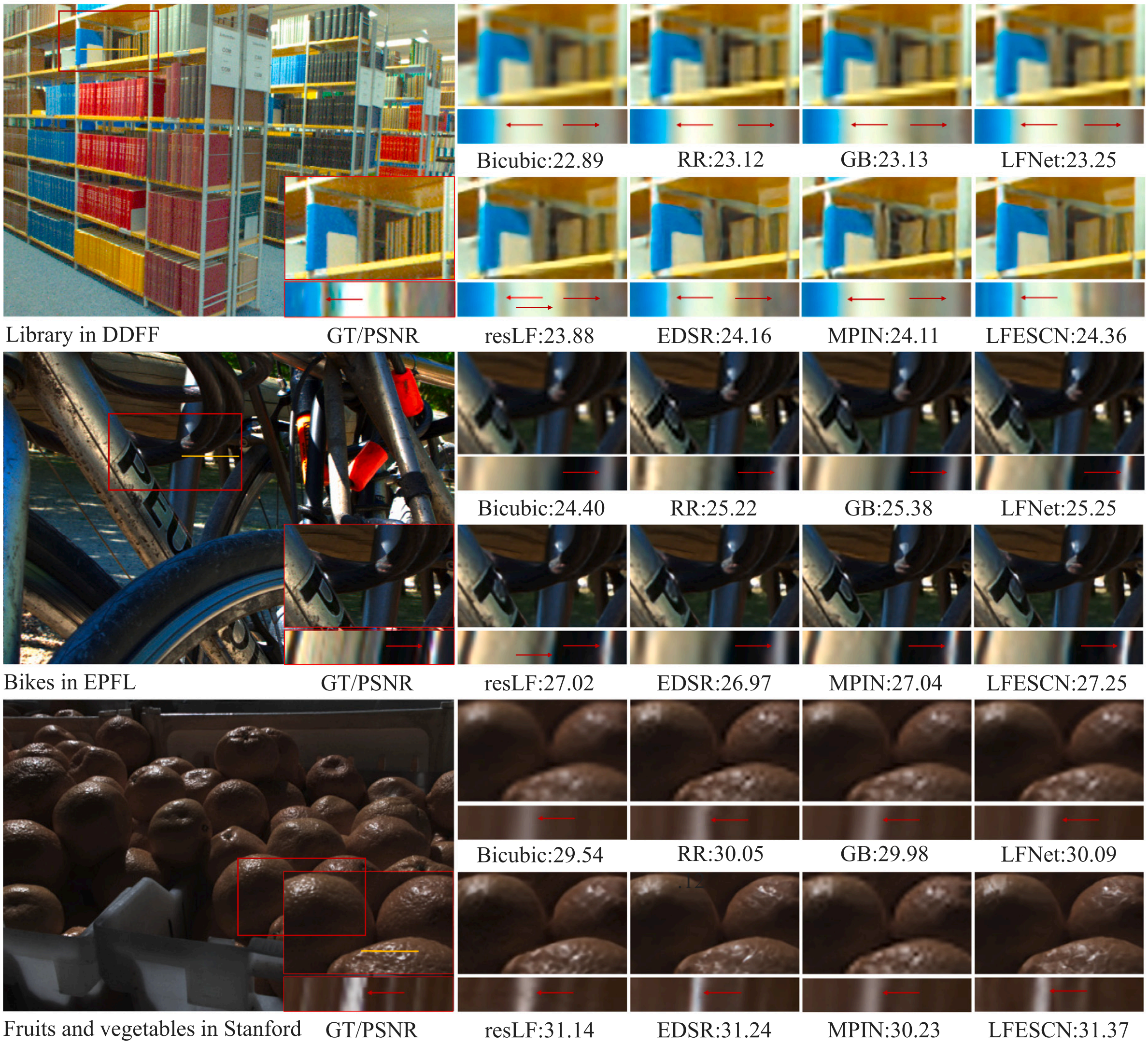


Fig. 8. The ×4 super-resolved central view images and corresponding EPIs of three real scene images are shown and the corresponding PSNR values of central views are illustrated below.

Table 6
Parameter numbers and testing time of several networks with leading LF SR performance.

Method	LFNet	EDSR	resLF	MPIN	LFESCN
Params (×10^6)	0.68	40.72	7.98	1.07	4.23
Time (s)	4.03	0.38	2.40	1.28	1.24

shift compensation, our method could make full use of the internal correlations to recover as much high-frequency information as possible.

4.7. Model analysis

Table 6 shows the number of parameters and running time of several networks with the leading LF SR performance. We test four methods on the whole dataset to calculate the average testing time of a single LF image on NVIDIA GTX 1080Ti for a fair comparison. As demonstrated in Table 6, our method is slower than the EDSR method in which all views can be processed in parallel. However, compared to resLF, due to the shallow network, our method is more efficient than resLF. Although our method has more parameters than the latest MPIN method, we can achieve better performance with the same running time.

5. Conclusion

In this paper, we have proposed a general epipolar shift compensation network for LF SR, called LFESCN. In our method, we employ multiple views as epipolar constraints to explore the internal relationships, which allow the entire light field with consistency across all sub-aperture images. Then, the sub-pixel shifts are learned from all epipolar directions through alignment by the compensation module, which would provide more high-frequency details for reconstruction. Especially, LFESCN is a general model for any LF images. The qualitative and quantitative results on publicly available datasets have demonstrated the superiority of our method over the state-of-the-art at different scale factors.

CRediT authorship contribution statement

Xinya Wang: Conceived and designed the research, Performed the experiments, Analyzed the results, Writing – original draft. **Jiayi Ma:** Conceived and designed the research, Provided insightful advices to this work and revised the manuscript. **Peng Yi:** Analyzed the results, Provided insightful advices to this work. **Xin Tian:** Conceived and designed the research, Provided insightful advices to this work and revised the manuscript. **Junjun Jiang:** Analyzed the results,

Provided insightful advices to this work and revised the manuscript. **Xiao-Ping Zhang:** Provided insightful advices to this work and revised the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was sponsored in part by the National Natural Science Foundation of China (61773295 and 61971315), in part by the Natural Science Foundation of Hubei Province (2019CFA037), and in part by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-04661).

References

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan, et al., Light field photography with a hand-held plenoptic camera, Comput. Sci. Tech. Rep. 2 (11) (2005) 1–11.

[2] A. Katayama, K. Tanaka, T. Oshino, H. Tamura, Dependent stereoscopic display using interpolation of multiviewpoint images, in: Stereoscopic Displays and Virtual Reality Systems II, 1995, pp. 11–20.

[3] J. Peng, Z. Xiong, D. Liu, X. Chen, Unsupervised depth estimation from light field using a convolutional neural network, in: Proceedings of the International Conference on 3D Vision, 2018, pp. 295–303.

[4] J. Peng, Z. Xiong, Y. Zhang, D. Liu, F. Wu, Lf-fusion: Dense and accurate 3D reconstruction from light field images, in: Proceedings of the IEEE Conference on Visual Communications and Image Processing, 2017, pp. 1–4.

[5] Z. Cheng, Z. Xiong, D. Liu, Light field super-resolution by jointly exploiting internal and external similarities, IEEE Trans. Circuits Syst. Video Technol. 30 (8) (2020) 2604–2616.

[6] T.E. Bishop, S. Zanetti, P. Favaro, Light field superresolution, in: Proceedings of the IEEE International Conference on Computational Photography, 2009, pp. 1–9.

[7] T.E. Bishop, P. Favaro, The light field camera: Extended depth of field, aliasing, and superresolution, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2011) 972–986.

[8] K. Mitra, A. Veeraraghavan, Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 22–28.

[9] M. Rossi, P. Frossard, Geometry-consistent light field super-resolution via graph-based regularization, IEEE Trans. Image Process. 27 (9) (2018) 4207–4218.

[10] M. Rossi, P. Frossard, Graph-based light field super-resolution, in: Proceedings of the IEEE International Workshop on Multimedia Signal Processing, 2017, pp. 1–6.

[11] J. Li, W. Guan, Adaptive lq-norm constrained general nonlocal self-similarity regularizer based sparse representation for single image super-resolution, Inf. Fusion 53 (2020) 88–102.

[12] H. Yin, S. Li, L. Fang, Simultaneous image fusion and super-resolution using sparse representation, Inf. Fusion 14 (3) (2013) 229–240.

[13] A. Panagiotopoulou, V. Anastassopoulos, Super-resolution image reconstruction techniques: Trade-offs between the data-fidelity and regularization terms, Inf. Fusion 13 (3) (2012) 185–195.

[14] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, T. Tan, Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution, IEEE Trans. Image Process. 27 (9) (2018) 4274–4286.

[15] Y. Yuan, Z. Cao, L. Su, Light-field image superresolution using a combined deep CNN based on EPI, IEEE Signal Process. Lett. 25 (9) (2018) 1359–1363.

[16] S. Zhang, Y. Lin, H. Sheng, Residual networks for light field image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11046–11055.

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

[18] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, Y. Guo, Spatial-angular interaction for light field image super-resolution, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 290–308.

[19] X. Wang, J. Ma, W. Gao, J. Jiang, Mpin: A macro-pixel integration network for light field super-resolution, Front. Inf. Technol. Electron. Eng. 22 (10) (2021) 1299–1310.

[20] S. Wanner, S. Meister, B. Goldluecke, Datasets and benchmarks for densely sampled 4D light fields, in: Proceedings of the Annual Workshop on Vision, Modeling and Visualization, 2013, pp. 225–226.

[21] R. Zhang, F. Nie, X. Li, X. Wei, Feature selection with multi-view data: A survey, Inf. Fusion 50 (2019) 158–167.

[22] Y. Du, W. Song, Q. He, D. Huang, A. Liotta, C. Su, Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection, Inf. Fusion 49 (2019) 89–99.

[23] R. Tsai, Multiple frame image restoration and registration, Adv. Comput. vis. Image Process 1 (1989) 1715–1989.

[24] M. Irani, S. Peleg, Improving resolution by image registration, CVGIP: Graph. Models Image Process. 53 (3) (1991) 231–239.

[25] D. Capel, A. Zisserman, Super-resolution from multiple views using learnt image models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 627–627.

[26] S. Farsiu, M.D. Robinson, M. Elad, P. Milanfar, Fast and robust multiframe super resolution, IEEE Trans. Image Process. 13 (10) (2004) 1327–1344.

[27] N. Nguyen, P. Milanfar, G. Golub, A computationally efficient superresolution image reconstruction algorithm, IEEE Trans. Image Process. 10 (4) (2001) 573–583.

[28] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.

[29] K.I. Kim, Y. Kwon, Single-image super-resolution using sparse regression and natural image prior, IEEE Trans. Pattern Anal. Mach. Intell. 32 (6) (2010) 1127–1133.

[30] M. Deudon, A. Kalaitzis, I. Goytom, M.R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S.E. Kahou, J. Cornebise, Y. Bengio, Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery, 2020, arXiv preprint arXiv:2002.06460.

[31] A.B. Molini, D. Valsesia, G. Fracastoro, E. Magli, Deepsum: Deep neural network for super-resolution of unregistered multitemporal images, IEEE Trans. Geosci. Remote Sens. 58 (5) (2019) 3644–3656.

[32] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, J. Nalepa, Deep learning for multiple-image super-resolution, IEEE Geosci. Remote Sens. Lett. 17 (6) (2019) 1062–1066.

[33] F. Salvetti, V. Mazzia, A. Khaliq, M. Chiaberge, Multi-image super resolution of remotely sensed images using residual attention deep neural networks, Remote Sens. 12 (14) (2020) 2207.

[34] J. Lim, H. Ok, B. Park, J. Kang, S. Lee, Improving the spatail resolution based on 4D light field data, in: Proceedings of the IEEE International Conference on Image Processing, 2009, pp. 1173–1176.

[35] F.P. Nava, J. Luke, Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera, in: Proceedings of the 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009, pp. 1–4.

[36] F. Pérez, A. Pérez, M. Rodríguez, E. Magdaleno, Fourier slice super-resolution in plenoptic cameras, in: Proceedings of the IEEE International Conference on Computational Photography, 2012, pp. 1–11.

[37] F. Pérez, A. Pérez, M. Rodríguez, E. Magdaleno, Super-resolved Fourier-slice refocusing in plenoptic cameras, J. Math. Imaging Vision 52 (2) (2015) 200–217.

[38] T.G. Georgiev, A. Lumsdaine, Super-resolution with the focused plenoptic camera, 2012, US Patent 8, 315, 476.

[39] C.-K. Liang, R. Ramamoorthi, A light transport framework for lenslet light field cameras, ACM Trans. Graph. 34 (2) (2015) 16.

[40] Y. Wang, G. Hou, Z. Sun, Z. Wang, T. Tan, A simple and robust super resolution method for light field images, in: Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 1459–1463.

[41] S. Wanner, B. Goldluecke, Variational light field analysis for disparity estimation and super-resolution, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2013) 606–619.

[42] M. Alain, A. Smolic, Light field super-resolution via LFBM5D sparse coding, in: Proceedings of the IEEE International Conference on Image Processing, 2018, pp. 2501–2505.

[43] R.A. Farrugia, C. Galea, C. Guillemot, Super resolution of light field images using linear subspace projection of patch-volumes, IEEE J. Sel. Top. Sign. Proces. 11 (7) (2017) 1058–1071.

[44] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, I. So Kweon, Learning a deep convolutional network for light-field image super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 24–32.

[45] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 184–199.

[46] H. Fan, D. Liu, Z. Xiong, F. Wu, Two-stage convolutional neural network for light field super-resolution, in: Proceedings of the IEEE International Conference on Image Processing, 2017, pp. 1167–1171.

[47] M.S.K. Gul, B.K. Gunturk, Spatial and angular resolution enhancement of light fields using convolutional neural networks, IEEE Trans. Image Process. 27 (5) (2018) 2146–2159.

[48] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.

[49] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets v2: More deformable, better results, 2018, arXiv preprint arXiv:1811.11168.

[50] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.

[51] N. Bodla, B. Singh, R. Chellappa, L.S. Davis, Soft-NMS–improving object detection with one line of code, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5561–5569.

[52] Y. Zhang, L. Shi, Y. Wu, K. Cheng, J. Cheng, H. Lu, Gesture recognition based on deep deformable 3D convolutional neural networks, Pattern Recognit. 107 (2020) 107416.

[53] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv preprint arXiv:1706.05587.

[54] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 466–481.

[55] Y. Tian, Y. Zhang, Y. Fu, C. Xu, Tdan: Temporally-deformable alignment network for video super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3360–3369.

[56] Z. Shi, X. Liu, K. Shi, L. Dai, J. Chen, Video frame interpolation via generalized deformable convolution, IEEE Trans. Multimed. (2021).

[57] H. Song, W. Xu, D. Liu, B. Liu, Q. Liu, D.N. Metaxas, Multi-stage feature fusion network for video super-resolution, IEEE Trans. Image Process. 30 (2021) 2923–2934.

[58] P. Yi, Z. Wang, K. Jiang, J. Jiang, T. Lu, J. Ma, A progressive fusion generative adversarial network for realistic and consistent video super-resolution, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

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

[60] X. Jia, B. De Brabandere, T. Tuytelaars, L.V. Gool, Dynamic filter networks, in: Advances in Neural Information Processing Systems, 2016, pp. 667–675.

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

[62] K. Honauer, O. Johannsen, D. Kondermann, B. Goldluecke, A dataset and evaluation methodology for depth estimation on 4d light fields, in: Proceedings of the Asian Conference on Computer Vision, 2016, pp. 19–34.

[63] M. Rerabek, T. Ebrahimi, New light field image dataset, in: Proceedings of the International Conference on Quality of Multimedia Experience, 2016, pp. 1–2.

[64] A.S. Raj, M. Lowney, R. Shah, G. Wetzstein, The stanford lytro light field archive, 2016, http://lightfields.stanford.edu/LF2016.html/ (Accessed: Oct. 22, 2018).

[65] C. Hazirbas, S.G. Soyer, M.C. Staab, L. Leal-Taixé, D. Cremers, Deep depth from focus, in: Proceedings of the Asian Conference on Computer Vision, 2018, pp. 525–541.