Spatial-Angular Interaction for Light Field Image Super-Resolution

Yingqian Wang\textsuperscript{1}, Longguang Wang\textsuperscript{1}, Jungang Yang\textsuperscript{1}, Wei An\textsuperscript{1}, Jingyi Yu\textsuperscript{2}, Yulan Guo\textsuperscript{1,3}

\textsuperscript{1}College of Electronic Science and Technology, National University of Defense Technology, China
\textsuperscript{2}School of Information Science and Technology, ShanghaiTech University, China
\textsuperscript{3}School of Electronics and Communication Engineering, Sun Yat-sen University, China
\{wangyingqian16, wanglongguang15, yangjungang, anwei, yulan.guo\}@nudt.edu.cn, yujingyi@shanghaitech.edu.cn

Abstract

Light field (LF) cameras record both intensity and directions of light rays, and capture scenes from a number of viewpoints. Both information within each perspective (i.e., spatial information) and among different perspectives (i.e., angular information) is beneficial to image super-resolution (SR). In this paper, we propose a spatial-angular interactive network (namely, LF-InterNet) for LF image SR. In our method, spatial and angular features are separately extracted from the input LF using two specifically designed convolutions. These extracted features are then repetitively interacted to incorporate both spatial and angular information. Finally, the interacted spatial and angular features are fused to super-resolve each sub-aperture image. Experiments on 6 public LF datasets have demonstrated the superiority of our method. As compared to existing LF and single image SR methods, our method can recover much more details, and achieves significant improvements over the state-of-the-arts in terms of PSNR and SSIM.

1. Introduction

Light field (LF) cameras provide multiple views of a scene, and thus enable many attractive applications such as post-capture refocusing [33], depth sensing [26], saliency detection [19], and de-occlusion [32]. However, LF cameras face a trade-off between spatial and angular resolutions [49]. That is, they either provide dense angular samplings with a low image resolution (e.g., Lytro\textsuperscript{1} and RayTrix\textsuperscript{2}), or capture high-resolution (HR) sub-aperture images (SAIs) with sparse angular samplings (e.g., camera arrays [37, 39]). Consequently, many efforts have been made to improve the angular resolution through LF reconstruction [39, 38], or the spatial resolution through LF image super-resolution (SR) [1, 46, 23, 31, 41]. In this paper, we focus on the LF image SR problem, namely, to reconstruct HR SAIs from their corresponding low-resolution (LR) SAIs.

Image SR is a long-standing problem in computer vision. To achieve high reconstruction performance, SR methods need to incorporate as much useful information as possible from LR inputs. In the area of single image SR, good performance can be achieved by fully exploiting the neighborhood context (i.e., spatial information) in an image. Using the spatial information, single image SR methods [5, 15, 20, 47] can successfully hallucinate missing details. In contrast, LF cameras capture scenes from multiple views. The complementary information among different views (i.e., angular information) can be used to further improve the performance of LF image SR.

However, due to the complicated 4D structures of LFs [18], it is highly challenging to incorporate spatial and angular information in an LF. Existing LF image SR methods

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Average PSNR and SSIM values achieved by state-of-the-art SR methods on 6 public LF datasets [22, 10, 56, 17, 29, 21]. Note that, our LF-InterNet improves PSNR and SSIM values by a large margin as compared to single image SR methods (VDSR [15], EDSR [20], RCAN [47]) and LF image SR methods (LFBM5D [11], resLF [45], GBSQ [23], LFSSR4D [41]).}
\end{figure}
fail to fully exploit both the angular information and the spatial information, resulting in limited SR performance. Specifically, in [43, 42, 44], SAI are first super-resolved separately using single image SR methods [5, 20], and then fine-tuned together to incorporate the angular information. The angular information is ignored by these two-stage methods [43, 42, 44] during their upsampling process. In [41, 46], only part of SAI are used to super-resolve one view, and the angular information in these discarded views is not incorporated. In contrast, Rossi et al. proposed a graph-based method [23] to consider all angular views in an optimization process. However, this method [23] cannot fully use the spatial information, and is inferior to deep learning-based SR methods [20, 47, 50, 41]. It is worth noting that, even all views are fed to a deep network, it is still challenging to achieve superior performance. Yeung et al. proposed a deep network named LF-SSR [41] to consider all views for LF image SR. However, as shown in Fig. 1 [20] LF-SSR [41] is inferior to resLF [46], EDSR [20], and RCAN [47].

The spatial information and the angular information are highly coupled in 4D LFs, and contribute to LF image SR in different manners. Consequently, it is difficult for networks to perform well directly using these coupled information. To efficiently incorporate spatial and angular information, we propose a spatial-angular interactive network (i.e., LF-InterNet) for LF image SR. We first specifically design two convolutions to decouple spatial and angular features from an input LF. Then, we develop LF-InterNet to repetitively interact and incorporate spatial and angular information. Extensive ablation studies have been conducted to validate our designs. We compare our method to the state-of-the-art single and LF image SR methods on 6 public LF datasets. As shown in Fig. 7, our LF-InterNet substantially improves the PSNR and SSIM performance as compared to existing SR methods.

2. Related Works

In this section, we review several major works on single image SR and LF image SR.

2.1. Single Image SR

In the area of single image SR, deep learning-based methods have been extensively explored. Readers are referred to recent surveys [34, 3, 40] for more details in single image SR. Here, we only review several milestone works. Dong et al. proposed the first CNN-based SR method (i.e., SRCNN [5]) by cascading 3 convolutional layers. Although SRCNN [5] is shallow and simple, it achieves significant improvements over traditional SR methods [28, 14, 45]. Afterwards, SR networks became increasingly deep and complex, and thus more powerful in spatial information exploitation. Kim et al. proposed a very deep SR network (i.e., VDSR [15]) with 20 convolutional layers. Global residual learning is applied to VDSR [15] to avoid slow convergence. Lim et al. proposed an enhanced deep SR network (i.e., EDSR [20]) with 65 convolutional layers by cascading residual blocks [9]. EDSR substantially improves its performance by applying both local and global residual learning, and won the NTIRE 2017 Challenge on single image SR [27]. More recently, Zhang et al. proposed a residual dense network (i.e., RDN [48]) with 149 convolutional layers by combining ResNet [9] with DenseNet [12]. Using residual dense connections, RDN [48] can fully extract hierarchical features for image SR, and thus achieve further improvements over EDSR [20]. Subsequently, Zhang et al. proposed a residual channel attention network (i.e., RCAN) [47] by applying both recursive residual mechanism and channel attention module [11]. RCAN [47] has 500 convolutional layers, and is one of the most powerful SR methods to date.

2.2. LF image SR

In the area of LF image SR, different paradigms have been proposed. Most early works follow the traditional paradigm. Bishop et al. [4] first estimated the scene depth and then used a de-convolution approach to estimate HR SAI. Wanner et al. [35] proposed a variational LF image SR framework using the estimated disparity map. Farrugia et al. [7] decomposed HR-LR patches into several subspaces, and achieved LF image SR via PCA analysis. Alain et al. extended SR-BM3D [6] to LFs, and super-resolved SAI using LFBM5D filtering [1]. Rossi et al. [23] formulated LF image SR as a graph optimization problem. These traditional methods [4, 35, 7, 23] use different approaches to exploit angular information, but cannot fully exploit spatial information.

In contrast, deep learning-based SR methods are more effective in exploiting spatial information, and thus can achieve promising performance. Many deep learning-based methods have been recently developed for LF image SR. In the pioneering work proposed by Yoon et al. (i.e., LFCNN [43]), SAI are first super-resolved separately using SRCNN [5], and then fine-tuned in pairs to incorporate angular information. Similarly, Yuan et al. proposed LF-DCNN [44], in which they used EDSR [20] to super-resolve each SAI and then fine-tuned the results. Both LFCNN [43] and LF-DCNN [44] handle the LF image SR problem in two stages, and do not use angular information in the first stage. Different from [43, 44], Wang et al. proposed LFNet [31] by extending BRCN [13] to LF image SR. In their method, SAI from the same row (or column) are fed to a recurrent network to incorporate the angular information. Zhang et al. stacked SAI along different angular directions to generate input volumes, and then proposed a multi-stream residual network named resLF [46]. Both LFNet [31] and resLF [46] reduce 4D LF to 3D LF by using part of SAI to super-
Note that, when an LF is organized as a 2D SAI array, the angular information will not be involved. In this way, the receptive field can be enlarged to cover pixels with disparities. Due to the 3D property of real scenes, objects of different depths have different disparity values in LFs. Consequently, pixels of an object among different views cannot always locate at a single macro-pixel. To handle this problem, we apply AFE and SFE for multiple times (i.e., performing spatial-angular interaction) in our network. As shown in Fig. 4, in this way, the receptive field can be enlarged to cover pixels with disparities.

3.2. Network Design

Our LF-InterNet takes an LR MPI of size $R^{AH \times AW}$ as its input and produces an HR SAI array of size $R^{\alpha AH \times \alpha AW}$, where $\alpha$ denotes the upsampling factor. Following [46, 41, 31, 44], we convert RGB images into an MPI representation in our method, and specifically design SSR [41] to alternately shuffle LF features between SAI pattern and macro-pixel image (MPI) pattern for convolution.
YCbCr color space, and only super-resolve Y channel images. An overview of our network is shown in Fig. 5.

3.2.1 Overall Architecture

Given an LR MPI $I_{LR} \in \mathbb{R}^{AH \times AW}$, the angular and spatial features are first extracted by AFE and SFE, respectively.

$$F_{A,0} = H_A(I_{LR}), \quad F_{S,0} = H_S(I_{LR}),$$

where $F_{A,0} \in \mathbb{R}^{H \times W \times C}$ and $F_{S,0} \in \mathbb{R}^{AH \times AW \times C}$ respectively represent the extracted angular and spatial features, $H_A$ and $H_S$ respectively represent the angular and spatial feature extractors (as described in Section 3.1). After initial feature extraction, features $F_{A,0}$ and $F_{S,0}$ are further processed by a series of interaction groups (i.e., Inter-Groups, see Section 3.2.2) to achieve spatial-angular feature interaction:

$$(F_{A,n}, F_{S,n}) = H_{IG,n}(F_{A,n-1}, F_{S,n-1}), \quad (n = 1, 2, \cdots, N),$$

where $H_{IG,n}$ denotes the $n^{th}$ Inter-Group and $N$ denotes the total number of Inter-Groups.

Inspired by RDN [48], we cascade all these Inter-Groups to fully use the information interacted at different stages. Specifically, features generated by each Inter-Group are concatenated and fed to a bottleneck block to fuse the interacted information. The feature generated by the bottleneck block is further added with the initial feature $F_{S,0}$ to achieve global residual learning. The fused feature $F_{S,t}$ can be obtained by

$$F_{S,t} = H_B([F_{A,t}, \cdots, F_{A,N}], [F_{S,1}, \cdots, F_{S,N}]) + F_{S,0},$$

where $H_B$ denotes the bottleneck block, $[]$ denotes the concatenation operation. Finally, the fused feature $F_{S,t}$ is fed to the reconstruction module, and an HR SAI array $I_{SR} \in \mathbb{R}^{AH \times AW}$ can be obtained by

$$I_{SR} = H_{1 \times 1}(S_{pix}(S_{lf}(H_S(F_{S,t})))�),$$

where $S_{lf}$, $S_{pix}$, and $H_{1 \times 1}$ represent LF shuffle, pixel shuffle, and $1 \times 1$ convolution, respectively. More details about feature fusion and reconstruction are introduced in Section 3.2.3.

3.2.2 Spatial-Angular Feature Interaction

The basic module for spatial-angular interaction is the interaction block (i.e., Inter-Block). As shown in Fig. 5(b), the Inter-Block takes a pair of angular and spatial features as inputs to achieve interaction. Specifically, the input angular feature is first upsampled by a factor of $A$. Here, a $1 \times 1$ convolution followed by a pixel shuffle layer is used for upsampling. Then, the upsampled angular feature is concatenated with the input spatial feature, and further fed to an SFE to incorporate the spatial and angular information. In this way, the complementary angular information can be used to guide spatial feature extraction. Simultaneously, the new angular feature is extracted from the input spatial feature by an AFE, and then concatenated with the input angular feature. The concatenated angular feature is further fed to a $1 \times 1$ convolution to integrate and update the angular information. Note that, the fused angular and spatial features are added with their respective input features to achieve local residual learning. In this paper, we cascade $K$ Inter-Blocks in an Inter-Group, i.e., the output of an Inter-Block forms the input of its subsequent Inter-Block. In summary, the spatial-angular feature interaction can be formulated as

$$F_{S,n}^{(k)} = H_S \left( \left[ F_{A,n}^{(k-1)}; F_{S,n}^{(k-1)} \right] \uparrow \right) + F_{S,n}^{(k-1)},$$

$$F_{A,n}^{(k)} = H_{1 \times 1} \left( \left[ F_{A,n}^{(k-1)}; H_A \left( F_{S,n}^{(k-1)} \right) \right] \right) + F_{A,n}^{(k-1)},$$

where $F_{S,n}^{(k)}$ and $F_{A,n}^{(k)}$ represent the output spatial and angular features of the $k^{th}$ Inter-Block in the $n^{th}$ Inter-Group, respectively, $\uparrow$ represents the upsampling operation.

3.2.3 Feature Fusion and Reconstruction

The objective of this stage is to fuse the interacted features to reconstruct an HR SAI array. The fusion and reconstruction stage mainly consists of bottleneck fusion (Fig. 5).
(c)), channel extension, LF shuffle (Fig. 5(d)), pixel shuffle (Fig. 5(e)), and final reconstruction.

In the bottleneck, the concatenated angular features \([\mathcal{F}_{A,1}, \ldots, \mathcal{F}_{A,N}] \in \mathbb{R}^{H \times W \times NC}\) are first fed to a \(1 \times 1\) convolution and a ReLU layer to generate a feature map \(\mathcal{F}_A \in \mathbb{R}^{H \times W \times C}\). Then, the squeezed angular feature \(\mathcal{F}_A\) is upsampled and concatenated with spatial features. The final fused feature \(\mathcal{F}_{S,t}\) can be obtained as

\[
\mathcal{F}_{S,t} = H_S ([\mathcal{F}_{S,1}, \ldots, \mathcal{F}_{S,N}, (\mathcal{F}_A \uparrow)] + \mathcal{F}_{S,0},
\]

After the bottleneck, we apply another SFE layer to extend the channel size of \(\mathcal{F}_{S,t}\) to \(\alpha^2C\) for pixel shuffle [25]. However, since \(\mathcal{F}_{S,t}\) is organized in the MPI pattern, we apply LF shuffle to convert \(\mathcal{F}_{S,t}\) into an SAI array representation for pixel shuffle. To achieve LF shuffle, we first extract pixels with the same angular coordinates in the MPI feature, and then re-organize these pixels according to their spatial coordinates, which can be formulated as

\[
\mathcal{I}_{\text{SAI}}(x, y) = \mathcal{I}_{\text{MPI}}(\xi, \eta),
\]

where

\[
x = H [(\xi/A) (1 - AH) + 1],
y = W [(\eta/A) (1 - AW) + 1].
\]

Here, \(x = 1, 2, \ldots, AH\) and \(y = 1, 2, \ldots, AW\) denote the pixel coordinates in the shuffled SAI arrays, \(\xi\) and \(\eta\) denote the corresponding coordinates in the input MPI, \([\cdot]\) represents the round-down operation. The derivation of Eqs. (7) and (8) is presented in the supplemental material.

Finally, a \(1 \times 1\) convolution is applied to squeeze the number of feature channels to 1 for HR SAI reconstruction.

4. Experiments

In this section, we first introduce the datasets and our implementation details, then conduct ablation studies to investigate our network. Finally, we compare our \(LF\text{-InterNet}\) to recent LF image SR and single image SR methods.

4.1. Datasets and Implementation Details

As listed in Table 1, we used 6 public LF datasets in our experiments. All the LFs in the training and test sets have an angular resolution of \(9 \times 9\). In the training stage, we first cropped each SAI into patches with a size of \(64 \times 64\), and then used bicubic downsampling with a factor of \(\alpha (\alpha = 2, 4)\) to generate LR patches. The generated LR patches were re-organized into MPI pattern to form the input of our network. The \(L_1\) loss function was used since it can generate good results for the SR task and is robust to outliers [2]. Following the recent works [46, 26], we augmented the training data by 8 times using random horizontal flipping, vertical flipping, and 90-degree rotation. Note that, during each data augmentation, all SAIs need to be flipped and rotated along both spatial and angular directions to maintain their LF structures.

By default, we used the model with \(N = 4, K = 4, C = 64\), and angular resolution of \(5 \times 5\) for both 2x and 4x SR.

![Figure 5. An overview of our LF-InterNet.](image)

Table 1. Datasets used in our experiments.

| Datasets | Type       | Training | Test |
|----------|------------|----------|------|
| EPFL [22]  | real-world | 70       | 10   |
| HCnew [10]  | synthetic   | 20       | 4    |
| HCold [36]  | synthetic   | 10       | 2    |
| INRIA [73]  | real-world | 35       | 5    |
| STFlytro [20] | real-world | 9        | 2    |
| STFlytro [21] | real-world | 250      | 50   |
| Total      | —          | 394      | 73   |

![Image of Table 1]
We also investigated the performance of other branches of our LF-InterNet in Section 4.2. We used PSNR and SSIM as quantitative metrics for performance evaluation. Note that, PSNR and SSIM were separately calculated on the Y channel of each SAI. To obtain the overall metric score for a dataset with $M$ scenes (each with an angular resolution of $A \times A$), we first obtain the score for a scene by averaging its $A^2$ scores, and then get the overall score by averaging the scores of all $M$ scenes.

Our LF-InterNet was implemented in PyTorch on a PC with an Nvidia RTX 2080Ti GPU. Our model was initialized using the Xavier method [8] and optimized using the Adam method [16]. The batch size was set to 12 and the learning rate was initially set to $5 \times 10^{-4}$ and decreased by a factor of 0.5 for every 10 epochs. The training was stopped after 40 epochs and took about 1 day.

### 4.2. Ablation Study

In this subsection, we compare the performance of our LF-InterNet with different architectures and angular resolutions to investigate the potential benefits introduced by different modules.

#### 4.2.1 Network Architecture

**Angular information.** We investigated the benefit of angular information by removing the angular path in LF-InterNet. That is, we only use SFE for LF image SR. Consequently, the network is identical to a single image SR network, and can only incorporate spatial information within each SAI. As shown in Table 2 only using the spatial information, the network (i.e., LF-InterNet-onlySpatial) achieves a PSNR of 29.75 and a SSIM of 0.893. Both the performance and the parameter number of LF-InterNet-onlySpatial is between VDSR [15] and EDSR [20].

**Spatial information.** To investigate the benefit introduced by spatial information, we changed the kernel size of all SFEs from $3 \times 3$ to $1 \times 1$. In this case, the spatial information cannot be exploited and integrated by convolutions. As shown in Table 2 the performance of LF-InterNet-onlyAngular is even inferior to that of bicubic interpolation.

| Model                        | PSNR  | SSIM  | Params. |
|------------------------------|-------|-------|---------|
| LF-InterNet-onlySpatial      | 29.75 | 0.893 | 1.27M   |
| LF-InterNet-onlyAngular      | 26.50 | 0.822 | 3.58M   |
| LF-InterNet-SACoupled        | 31.02 | 0.918 | 5.10M   |
| LF-InterNet                 | 31.65 | 0.925 | 5.23M   |

| Method       | PSNR  | SSIM  | Params. |
|--------------|-------|-------|---------|
| Bicubic      | 27.84 | 0.855 |         |
| VDSR [15]    | 29.17 | 0.880 | 0.66M   |
| EDSR [20]    | 30.29 | 0.903 | 1.45M   |

That is because, neighborhood context in an image is highly significant in recovering details. Consequently, spatial information plays a major role in LF image SR, while angular information can only be used as a complimentary part to spatial information but cannot be used alone.

**Information decoupling.** To investigate the benefit of spatial-angular information decoupling, we stacked all SAIs along the channel dimension as input, and used $3 \times 3$ convolutions with a stride of 1 to extract both spatial and angular information from these stacked images. Note that, the cascaded framework with global and local residual learning was maintained to keep the overall network architecture unchanged, and the feature depth was set to 128 to keep the number of parameters comparable to that of LF-InterNet. As shown in Table 3 LF-InterNet-SACoupled is inferior to LF-InterNet. That is, with comparable number of parameters, LF-InterNet can handle the 4D LF structure and achieve LF image SR in a more efficient way.

| AngRes | Scale | PSNR  | SSIM  | Params. |
|--------|-------|-------|-------|---------|
| 3 $\times$ 3 | $\times$ 2 | 37.95 | 0.980 | 2.73M   |
| 5 $\times$ 5 | $\times$ 2 | 38.81 | 0.983 | 4.80M   |
| 7 $\times$ 7 | $\times$ 2 | 39.05 | 0.984 | 7.90M   |
| 9 $\times$ 9 | $\times$ 2 | 39.08 | 0.985 | 12.02M  |
| 3 $\times$ 3 | $\times$ 4 | 31.30 | 0.918 | 3.15M   |
| 5 $\times$ 5 | $\times$ 4 | 31.84 | 0.927 | 5.23M   |
| 7 $\times$ 7 | $\times$ 4 | 32.04 | 0.931 | 8.33M   |
| 9 $\times$ 9 | $\times$ 4 | 32.07 | 0.933 | 12.48M  |

Table 2. Comparative results achieved on the STFlytro dataset [21] by our LF-InterNet with different settings for $4 \times$ SR. Note that, the results of bicubic interpolation, VDSR [15], and EDSR [20] are also listed as baselines.

Table 3. Comparative results achieved on the STFlytro dataset [21] by our LF-InterNet with different number of interactions for $4 \times$ SR.

Table 4. Comparative results achieved on the STFlytro dataset [21] by our LF-InterNet with different angular resolutions for $2 \times$ and $4 \times$ SR.
As the number of interactions increases, the performance is steadily improved. This clearly demonstrates the effectiveness of our spatial-angular feature interaction mechanism.

4.2.2 Angular Resolution

We compared the performance of LF-InterNet with different angular resolutions. Specifically, we extracted the central $A \times A (A = 3, 5, 7, 9)$ SAIs from the input LFs, and trained different models for both $2 \times$ and $4 \times$ SR. As shown in Table 4, the PSNR and SSIM values for both $2 \times$ and $4 \times$ SR are improved as the angular resolution is increased. That is because, additional views provide rich angular information for LF image SR. It is also notable that, the improvements tend to be saturated as the angular resolution increases from $7 \times 7$ to $9 \times 9$ (only 0.03 dB improvement in PSNR). That is because, the complementary information provided by additional views is already sufficient. As the angular information is fully exploited, a further increase of views can only provide minor performance improvements.

4.3. Comparison to the State-of-the-arts

We compare our method to three milestone single image SR methods (i.e., VDSR [15], EDSR [20], and RCAN [47]) and four state-of-the-art LF image SR methods (i.e., LFBSM5D [1], GBSQ [23], LFSSR [41], and resLF [46]). All these methods were implemented using their released codes and pre-trained models. We also present the results of bicubic interpolation as the baseline results. For simplification, we only present the results on $5 \times 5$ LFs for $2 \times$ and $4 \times$ SR. Since the angular resolution of LFSSR [41] is fixed, we use its original version with $8 \times 8$ input SAIs.

Quantitative Results. Quantitative results are presented in Table 5. For both $2 \times$ and $4 \times$ SR, our method (i.e., LF-InterNet$_{64}$) achieves the best results on all the 6 datasets and surpasses existing methods by a large margin. For example, 2.04 dB and 1.51 dB PSNR improvements in average over the state-of-the-art LF image SR method resLF [46] can be observed for $2 \times$ and $4 \times$ SR, respectively. It is worth noting that, even the feature depth of our model is halved to 32, our method (i.e., LF-InterNet$_{32}$) can still achieve the highest SSIM scores on all the 6 datasets and the highest PSNR scores on 5 of the 6 datasets as compared to existing methods. Note that, the numbers of parameters of LF-InterNet$_{32}$ are only 1.20M for $2 \times$ SR and 1.31M for $4 \times$ SR, which are significantly smaller than recent deep learning-based SR methods [47, 41, 46].

Qualitative Results. Qualitative results of $2 \times$ and $4 \times$ SR are shown in Figs. 6 and 7 with more visual comparisons being provided in our supplemental material. It can be observed from Fig. 6 that, our method can well preserve the textures and details (e.g., the horizontal stripes in the scene HClnew_origami and the stairway in the scene INRIA_Sculpture) in these super-resolved images. In contrast, although the single image SR method RCAN [47] achieves high PSNR and SSIM scores, the images generated by RCAN [47] are over-smoothed and poor in details. It can be observed from Fig. 7 that the visual superiority of our method is more obvious for $4 \times$ SR. Since the input LR images are severely degraded by the down-sampling
operation, the process of 4×SR is highly ill-posed. Single image SR methods use spatial information only to hallucinate missing details, and they usually generate ambiguous and even fake textures (e.g., the window frame in scene EPFL_Palais generated by RCAN [47]). In contrast, LF image SR methods can use complementary angular information among different views to produce authentic results. However, the results generated by existing LF image SR methods [23, 41, 46] are relatively blurring. As compared to these single image and LF image SR methods, the results produced by our LF-InterNet are much more close to the groundtruth images.

**Performance w.r.t. Perspectives.** Since our LF-InterNet can super-resolve all SAIs in an LF, we further investigate the reconstruction quality with respect to different perspectives. We used the central 7 × 7 views of scene HCIold_MonasRoom [36] as input to perform both 2× and 4× SR. The PSNR and SSIM values are calculated for each perspective and are visualized in Fig. 8. Since resLF [46] uses part of views to super-resolve different perspectives, the reconstruction qualities of resLF [46] for non-central views are relatively low. In contrast, our LF-InterNet jointly uses the angular information from all input views to super-resolve each perspective, and thus achieves much higher reconstruction qualities with a more balanced distribution among different perspectives.

Figure 6. Visual results of 2× SR.

Figure 7. Visual results of 4× SR.

Figure 8. Visualizations of PSNR and SSIM values achieved by resLF [46] and LF-InterNet on each perspective of scene HCIold_MonasRoom [36]. Here, 7×7 input views are used to perform both 2× and 4× SR. Our LF-InterNet achieves high reconstruction qualities with a balanced distribution among different SAIs.
5. Conclusion

In this paper, we proposed a deep convolutional network LF-InterNet for LF image SR. We first introduce an approach to extract and decouple spatial and angular features, and then design a feature interaction mechanism to incorporate spatial and angular information. Experimental results have clearly demonstrated the superiority of our method. Our LF-InterNet outperforms the state-of-the-art SR methods by a large margin in terms of PSNR and SSIM, and can recover rich details in the reconstructed images.

6. Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (No. 61972435, 61602499), Natural Science Foundation of Guangdong Province, Fundamental Research Funds for the Central Universities (No. 18lgzd06).

References

[1] Martin Alain and Aljosa Smolic. Light field super-resolution via lfbm5d sparse coding. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 2501–2505. IEEE, 2018.

[2] Yildiray Anagun, Sahin Isik, and Erol Seke. Srlibrary: Comparing different loss functions for super-resolution over various convolutional architectures. Journal of Visual Communication and Image Representation, 61:178–187, 2019.

[3] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. arXiv preprint arXiv:1904.07523, 2019.

[4] Tom E Bishop and Paolo Favaro. The light field camera: Extended depth of field, aliasing, and super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(5):972–986, 2011.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In European Conference on Computer Vision (ECCV), pages 184–199. Springer, 2014.

[6] Karen Egiazarian and Vladimir Katkovnik. Single image super-resolution via bm3d sparse coding. In European Signal Processing Conference (EUSIPCO), pages 2849–2853. IEEE, 2015.

[7] Reuben A Farrugia, Christian Galea, and Christine Guillemot. Super resolution of light field images using linear subspace projection of patch-volumes. IEEE Journal of Selected Topics in Signal Processing, 11(7):1058–1071, 2017.

[8] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics, pages 249–256, 2010.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[10] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In Asian Conference on Computer Vision (ACCV), pages 19–34. Springer, 2016.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7132–7141, 2018.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4700–4708, 2017.

[13] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In Advances in Neural Information Processing Systems (NeurIPS), pages 235–243, 2015.

[14] Yang Jiahao, Wright John, Huang Thomas, and Ma Yi. Image super-resolution via sparse representation. IEEE Transactions on Image Processing, 19(11):2861–2873, 2010.

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1646–1654, 2016.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning and Representation (ICLR), 2015.

[17] Mikael Le Pendum, Xiaoran Jiang, and Christine Guillemot. Light field inpainting propagation via low rank matrix completion. IEEE Transactions on Image Processing, 27(4):1981–1993, 2018.

[18] Marc Levoy and Pat Hanrahan. Light field rendering. In Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pages 31–42. ACM, 1996.

[19] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2806–2813, 2014.

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 136–144, 2017.

[21] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. Stanford lytro light field archive, 2016.

[22] Martin Rerabek and Touradj Ebrahimi. New light field image super-resolution via graph-based regularization. IEEE Transactions on Image Processing, 27(9):4207–4218, 2018.

[23] Mattia Rossi and Pascal Frossard. Geometry-consistent light field super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4700–4708, 2017.

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via attention.
gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[25] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan 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 (CVPR)*, pages 1874–1883, 2016.

[26] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epine: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4748–4757, 2018.

[27] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 114–125, 2017.

[39] Gaohuang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field reconstruction using deep convolutional network on epi. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6319–6327, 2017.

[40] Wenming Yang, Xuechen Zhang, Yaping Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 2019.

[41] Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018.

[42] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24(6):848–852, 2017.

[43] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 24–32, 2015.

[44] Yan Yuan, Ziqi Cao, and Lijuan Su. Light-field image super-resolution using a combined deep cnn based on epi. *IEEE Signal Processing Letters*, 25(9):1359–1363, 2018.

[45] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on Curves and Surfaces*, pages 711–730. Springer, 2010.

[46] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light-field image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11046–11055, 2019.

[47] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[48] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.

[49] Hao Zhu, Mantang Guo, Hongdong Li, Qing Wang, and Antonio Robles-Kelly. Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution. *IEEE transactions on visualization and computer graphics*, 2019.
Spatial-Angular Interaction for Light Field Image Super-Resolution

Supplemental Material

Section A presents details of light field (LF) shuffle. Section B provides additional visual comparisons.

A. Light Field Shuffle

We use the notations in Table 6 for formulation. As shown in Fig. 9, an LF \( \mathcal{L} \in \mathbb{R}^{U \times V \times H \times W} \) can be organized into a macro-pixel image \( \mathcal{I}_{\text{MPI}} \in \mathbb{R}^{UH \times VW} \) or an array of sub-aperture images \( \mathcal{I}_{\text{SAIs}} \in \mathbb{R}^{UH \times VW} \). Consequently, LF shuffle is defined as the transformation between these two representations. To convert LFs from one representation to the other representation, the one-to-one mapping function between MPI and SAIs needs to be built. Without loss of generality, we take spatial-to-angular shuffle as an example, namely, to find point \((\xi, \eta)\) in \(\mathcal{I}_{\text{MPI}}\) corresponding to a known point \((x, y)\) in \(\mathcal{I}_{\text{SAIs}}\). We first calculate the angular coordinates \(u\) and \(v\) of point \((x, y)\) according to

\[
\begin{align*}
  u &= \lfloor x/H \rfloor + 1, \\
  v &= \lfloor y/W \rfloor + 1.
\end{align*}
\]  

(9)

Using the angular coordinates, the spatial coordinates \(h\) and \(w\) can be derived by

\[
\begin{align*}
  h &= x - (u - 1) \cdot H = x - \lfloor x/H \rfloor \cdot H, \\
  w &= y - (v - 1) \cdot W = y - \lfloor y/W \rfloor \cdot W.
\end{align*}
\]  

(10)

Since \(\mathcal{I}_{\text{SAIs}}\) and \(\mathcal{I}_{\text{MPI}}\) represent the same LF, \((x, y)\) and \((\xi, \eta)\) in these two representations have the same spatial and angular coordinates. Therefore, we find \((\xi, \eta)\) corresponding to \((u, v, h, w)\) as follows:

\[
\begin{align*}
  \xi &= U \cdot (h - 1) + u \\
  &= U \cdot (x - \lfloor x/H \rfloor \cdot H - 1 + \lfloor x/H \rfloor + 1) + 1 \\
  &= U \cdot (x - 1) \cdot (1 - U \cdot H) + 1 \\
  \eta &= V \cdot (w - 1) + v \\
  &= V \cdot (y - \lfloor y/W \rfloor \cdot W - 1 + \lfloor y/W \rfloor + 1) + 1 \\
  &= V \cdot (y - 1) \cdot (1 - V \cdot W) + 1
\end{align*}
\]  

(11)

(12)

The angular-to-spatial shuffle can be derived following a similar approach. That is,

\[
\begin{align*}
  x &= H \cdot (u - 1) + h \\
  &= H \cdot (\xi - 1) + \lfloor \xi/U \rfloor \cdot (1 - U \cdot H) + 1
\end{align*}
\]  

(13)

\[
\begin{align*}
  y &= W \cdot (v - 1) + w \\
  &= W \cdot (\eta - 1) + \lfloor \eta/V \rfloor \cdot (1 - V \cdot W) + 1
\end{align*}
\]  

(14)

Note that, Eq. (8) in the main body of our manuscript can be derived from the above equations (i.e., Eqs. (13-14)) by assigning \(A\) to \(U\) and \(V\).

B. Additional Visual Comparisons

Additional visual comparisons for 2× and 4× SR are shown in Figs. 10 and 11, respectively.

Table 6. Notations used in this supplemental material.

| Notation            | Representation               |
|---------------------|-----------------------------|
| \(\mathcal{L} \in \mathbb{R}^{U \times V \times H \times W}\) | a 4D LF                     |
| \(\mathcal{I}_{\text{SAIs}} \in \mathbb{R}^{UH \times VW}\) | a 2D SAI array              |
| \(\mathcal{I}_{\text{MPI}} \in \mathbb{R}^{UH \times VW}\) | a 2D MPI                    |
| \(U, V \in \mathbb{Z}_{+}\) | angular size                |
| \(h, W \in \mathbb{Z}_{+}\) | spatial size                |
| \(u, v \in \mathbb{Z}_{+}\) | angular coordinate          |
| \(h, w \in \mathbb{Z}_{+}\) | spatial coordinate          |
| \((x, y) \in \mathbb{Z}_{+}^2\) | coordinate in \(\mathcal{I}_{\text{SAIs}}\) |
| \((\xi, \eta) \in \mathbb{Z}_{+}^2\) | coordinate in \(\mathcal{I}_{\text{MPI}}\) |
| \(\lfloor \cdot \rfloor\) | round-down operation         |

Figure 9. An illustration of LF shuffle. Since the SAIs and the MPI denote the same LF, the objective of LF shuffle is to re-organize LFs between these two representations.
| Dataset                  | Groundtruth | Bicubic | RCAN | GBSQ | LFSSR_4D | resLF | Ours | PSNR/SSIM |
|-------------------------|-------------|---------|------|------|----------|-------|------|-----------|
| EPFL luxembourg          |             |         |      |      |          |       |      | 37.46/0.975 |
| HCIold_buddha            |             |         |      |      |          |       |      | 37.46/0.975 |
| HCnew_bicycle            |             |         |      |      |          |       |      | 37.46/0.975 |
| INRIA_messydesk          |             |         |      |      |          |       |      | 37.46/0.975 |
| STFlytro_cards           |             |         |      |      |          |       |      | 37.46/0.975 |
| STFlytro_cars_52         |             |         |      |      |          |       |      | 37.46/0.975 |
| STFlytro_general_23      |             |         |      |      |          |       |      | 37.46/0.975 |
| STFlytro_people_3        |             |         |      |      |          |       |      | 37.46/0.975 |

Figure 10. Additional visual results for 2×SR.
| Image                  | Groundtruth | Bicubic | RCAN    | GBSQ    | LFSSR_4D | resLF | Ours     |
|------------------------|-------------|---------|---------|---------|----------|-------|----------|
| EPFL_friends_1         | 27.69/0.931 | 31.26/0.959 | 28.16/0.943 | 28.84/0.958 | 29.69/0.959 |       | 32.26/0.966 |
| HCInew_bedroom         | 27.15/0.831 | 29.26/0.875 | 28.28/0.864 | 28.97/0.876 | 29.71/0.889 |       | 30.47/0.901 |
| HCIold_monai           | 32.04/0.934 | 36.60/0.965 | 33.11/0.950 | 35.39/0.965 | 36.25/0.969 |       | 37.79/0.977 |
| INRIA_habais           | 24.84/0.850 | 28.63/0.914 | 26.20/0.890 | 27.32/0.912 | 28.20/0.924 |       | 31.03/0.942 |
| STFlytro_cards         | 23.60/0.778 | 26.83/0.897 | 25.61/0.859 | 26.04/0.872 | 26.76/0.895 |       | 27.67/0.917 |
| STFlytro_cars_40       | 29.08/0.914 | 32.08/0.951 | 29.53/0.923 | 31.98/0.952 | 30.66/0.938 |       | 33.61/0.963 |
| STFlytro_flowers_29    | 27.31/0.867 | 30.09/0.914 | 27.97/0.884 | 30.34/0.922 | 29.17/0.903 |       | 32.42/0.941 |
| STFlytro_reflective_21 | 30.52/0.906 | 32.31/0.932 | 30.72/0.912 | 32.04/0.930 | 31.24/0.921 |       | 32.59/0.935 |

Figure 11. Additional visual results for 4×SR.