Light Field Super-Resolution Using Complementary-View Feature Attention

Wei Zhang[1], Wei Ke[1], Da Yang[2,3], Hao Sheng[1,2,3] and Zhang Xiong[1,2,3]
[1] Faculty of Applied Sciences, Macao Polytechnic University,
Macao SAR 999078, P.R. China
[2] State Key Laboratory of Software Development Environment,
School of Computer Science and Engineering, Beihang University,
Beijing 100191, P.R. China
[3] Beihang Hangzhou Innovation Institute Yuhang,
Xixi Octagon City, Yuhang District, Hangzhou 310023, P.R. China
{wei.zhang, wke}@ipm.edu.mo, {da.yang, shenghao, xiongz}@buaa.edu.cn

Abstract

Light field (LF) cameras record multiple perspectives through a sparse sampling of real scenes, and these perspectives provide useful information for each other. This information is beneficial to the LF super-resolution (LFSR). Comparing with traditional single-image super-resolution (SISR), LF has the parallax structure and perspective correlation between LF views. Furthermore, the performance of existing methods is limited as they fail to deeply explore the complementary information across LF views. In this paper, we propose a novel network, called light field complementary-view feature attention network (LF-CFANet), to improve LFSR by dynamically learning the complementary information among LF views. Specifically, we design a residual complementary-view spatial and channel attention module (RCSCAM) to effectively interact complementary information between complementary views. Moreover, RCSCAM captures the relationship of different channels, and is able to generate informative features for reconstructing LF images while ignoring the redundant information. Then, a maximum-difference information supplementary branch (MDISB) is used to supplement information from maximum-difference angular positions based on the geometrical structure of LF images. MDISB can guide the process of reconstruction. Experimental results on both synthetic and real-world datasets demonstrate the superiority of our method. The proposed LF-CFANet has a more advanced reconstruction performance that displays faithful details with better SR accuracy than state-of-the-art methods.

1. Introduction

Light field (LF) provide 4D LF images compared with conventional cameras, and thus LF imaging technology has been used widely in many applications, such as VR [1, 2], 3D reconstruction [3, 4], saliency detection [5, 6] and post-capture image editing [7]. One of the most popular applications is LF cameras, e.g., Lytro and RayTrix. As illustrated in Fig. 1(a), these cameras place a micro-lens array between the main lens and the sensor to provide multiple views for a scene, which are different from conventional cameras. LF images, captured by the handheld LF camera [1, 2], record the spatial information (accumulation from the same object point) and the angular information (the intensity values by all ray directions). However, due to the limitation of sensor resolution, the spatial resolution of LF images is much lower than that of commercial 2D cameras. Therefore, image super-resolution (SR) technology plays an important role in LF applications, and this technology effectively promotes the field of LF.

LF super-resolution (LFSR) is an ill-posed problem. This problem can be solved by exploring the efficient use of sub-pixel information from different views to reconstruct SR images. Traditional methods generally solve the SR problem among multiple views based on prior disparity information, such as Bayesian framework [8], variational framework [9, 10] and Gaussian mixture framework [11]. However, these methods are restricted by the inaccurate prior disparity information, and their computational costs are very high. With the development of deep learning, learning-based methods [12–15] are used to address the problem of complex 4D structure of LF data, and improve the performance compared to the traditional approaches. Although continuous improvements have been investigated [16, 17], the inherent complementary information among sub-aperture images (SAI) still fails to be

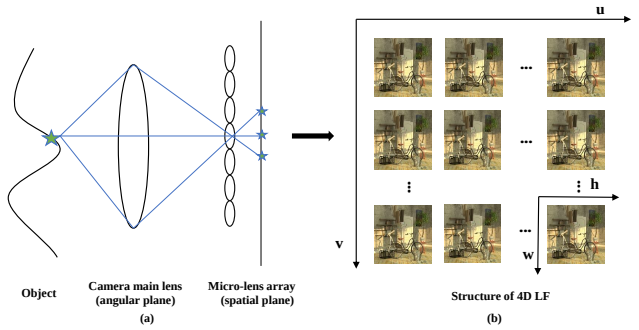


Figure 1. Principle of LF camera and structure of 4D LF. (a) Illustration of the schematic of LF cameras. (b) An example of 4D LF structure. The angular position (u, v) of an LF is determined by the number of sensor pixels under each micro-lens, while the spatial position (h, w) is related to the number of micro-lenses in the array.

fully explored, because the parallax information is treated equally for each view, and the feature fusion between complementary views is not sufficient. These issues hinder the performance improvement of LFSR methods.

Taking advantage of the attention mechanism in SR networks [18–20], we propose a spatial and channel attention network, namely light field complementary-view feature attention network (LF-CFANet), to improve the spatial resolution of LF images. As shown in Fig. 2, this network consists of two main modules, namely the residual complementary-view spatial and channel attention module (RCSCAM) and the maximum-difference information supplementary branch (MDISB). Specifically, RCSCAM is designed to fuse the complementary information among the pairs of LF images. With our RCSCAM, the reconstruction features can be interacted with the complementary sub-pixel information and local similarity information from different auxiliary views by computing an attention map. Meanwhile, this module with a channel attention mechanism can capture the global channel-level information by adaptively adjusting the response value of the feature map of each channel. To guide LF reconstruction both effectively and efficiently, we propose the MDISB to guide the features for SR reconstruction and obtain the maximum-difference information among LF views. In MDISB, the features of a reference view and four auxiliary views are collected from reservoir based on the maximum-difference angular positions. The maximum-difference feature is used as a guide for the reconstruction of the reference view. Through these two modules, the complementary information in the across LF views can be effectively utilized to reconstruct the SR LF images to a certain extent.

Extensive experimental results over the LF datasets (real-world and synthetic) demonstrate that the proposed method achieves both higher quantitative and better qualitative performance, compared with the state-of-the-art methods. Our contributions are summarized as follows.

1. We propose an RCSCAM to better exploit correlation cues for LF complementary-view pairs and generate the effectively fused complementary-view features by introducing the attention mechanism. The channel attention increases the global perception of feature channels, and the spatial attention mechanism enhances the interaction of spatial information between complementary views.

2. We develop an MDISB for guiding supplementation with the most difference information for SR views by treating each perspective unequally. The information can be provided form reservoir by concatenating two feature pairs consisting of four maximum-difference fused features based on the parallax structure of the LF images.

3. Our LF-CFANet explores the effectiveness of using the attention mechanism for feature interaction in LF complementary views. Extensive experiments have demonstrated the performance improvements compared with the state-of-the-art methods.

The rest of this paper is organized in the following sections. Section 2 introduces a brief review of the related work. The structure of LF and the architecture of our LF-CFANet are outlined in Section 3. In Section 4, we give extensive analysis and experiments by using synthetic and real-world datasets. Finally, Section 5 summarizes the conclusion of this paper.

2. Related work

In this section, we review the related work on both single image super-resolution (SISR) and LFSR.

2.1. Single image super-resolution

SISR is a reconstruction technology for fuzzy low-resolution (LR) images. This technology plays an important role in the field of surveillance, satellite imaging, microscopic imaging, etc. Several studies [21, 22] provided more details in reviewing SISR. Here, we give a review of several recent advancements. Nowadays, deep learning has gradually become a research hot spot, and has great influence on the technology of super-resolution. Dong et al. [23, 24] proposed a milestone study in an SR deep convolutional neural network (SRCNN), a seminal method in the field of SR. This simple and shallow model shows better reconstruction quality than earlier work. Additionally, Kim et al. [25] proposed a very deep convolutional network (VDSR) combined with residual learning, which was more efficient and

achieved higher quality over that of Dong's [23, 24]. Especially, VDSR could obtain a larger receptive field by stacking filters, the problem of slow convergence was solved by applying global residual learning. To make good use of intra-view information, more powerful models have appeared based on deep networks. Lim et al. [26] proposed an enhanced deep SR network (EDSR). This network achieved extraordinarily well performance than previous methods by revising the residual module and multi-scale model [27]. Zhang et al. [28, 29] proposed a residual dense network (RDN), which could make full use of all hierarchical features in all convolutional layers and provided better performance in feature extraction than EDSR. With the application of the attention mechanism, Zhang et al. [30] proposed a residual channel attention network (RCAN), that worked by inserting a channel attention module for considering the interdependence between channels. Recently, Dai et al. [31] proposed a second-order attention network (SAN) by applying the trainable second-order attention module to capture spatial information. Both the RCAN and SAN have achieved promising performance in SISR reconstruction.

As shown in the above review, SISR methods efficiently and effectively reconstruct the spatial information of single images. However, these methods cannot directly handle the correlation among multiple views, they cannot be applied to the field of LFSR.

2.2. LF super-resolution

For LFSR, a straightforward way is fine-tuning the network parameters of SISR. However, LFSR is more focused on complementary information, which are provided by multiple LF images from one scene to reconstruct an SR image. Existing LFSR methods can be mainly divided into two categories: optimization-based and learning-based approaches.

Optimization-based approaches reconstruct SR images based on the estimated disparities among different views. Bishop and Favaro [8] first utilized a Bayesian framework for LFSR. Wanner and Goldluecke [9, 10] proposed a variational method for SR by introducing the disparity maps obtained from EPIs. Mitra and Veeraraghavan [11] proposed a patch based approach modeled by a Gaussian mixture model to solve the LF problems. The framework of this method could handle many different processing tasks. To better supplement complementary information and avoid costly disparity estimation, Rossi and Frossard [32] proposed an LFSR framework for the homogeneous reconstruction of all views in the LF by using a graph-based regularizer. After this, Alain and Smolic [33] have proposed a method to convert the inverse problem of LFSR into an optimization problem based on prior sparsity. Although these methods could well encode the complex 4D LF, optimization-based methods were not sufficient to supplement the spatial information among different views.

Learning-based approaches demonstrate superiority to optimization-based approaches in using complementary information among different views. Making full use of complementary information can improve the quality of LFSR. Yoon et al. [13, 14] have proposed a pioneering work introducing CNN to the field of LF (LFCNN), while Yuan et al. [12] proposed an SR method that fully exploited the particular structure of the LF with an SISR module and an EPI enhancement module. These modules well maintained the structural characteristics of LF. By extending BRCN [34], Wang et al. [35] proposed a bidirectional recurrent convolutional neural network (namely, LFNet) and stacked generalization techniques to synthesize the final sub-aperture images. In this structure, the recurrent neural network was improved to handle the structure of horizontal and vertical directions. Within the network, the spatial correlations between neighboring views could be modified to be more effective and flexible. Inspired by residual network, Zhang et al. [16] proposed a multi-branch residual network (resLF) which handled image stacks with consistent sub-pixel offsets, and each branch could extract high-frequency details from LF images. In addition, in order to preserve the parallax structure, Jin et al. [36] proposed a method with a two-step LF spatial resolution by introducing a perspective feature fusion module and the structural consistency regularization loss (LF-ATO). More recently, Wang et al. [37] proposed an LF-InterNet to extract and incorporate spatial and angular information. This network could gradually interact the spatial and angular information. The result of this method has achieved high accuracy for LF reconstruction.

In summary, these methods implicitly learn the internal correspondence of the LF structure, and they are gradually making improvement to LFSR. However, due to the design of the network structure, complementary information is still not fully utilized. For example, LFNet designs a bidirectional recurrent network to fuse angular information among SAIs. This information only contains row and column directions, and it cannot be efficiently used to reconstruct LF images. Consequently, we propose a complementary-view feature attention approach using the information of all auxiliary views to reconstruct the reference view.

3. Architecture of LF-CFANet

In this section, we introduce the 4D LF representation, and propose a many-to-one LFSR network. The architecture of our LF-CFANet is shown in Fig. 2. It is noteworthy that the part of feature fusion is composed of two branches (MDISB and a reservoir branch). For the target input of our network, we convert LF images from RGB color space to YCbCr color space and only super-resolve the Y-channel images [36].

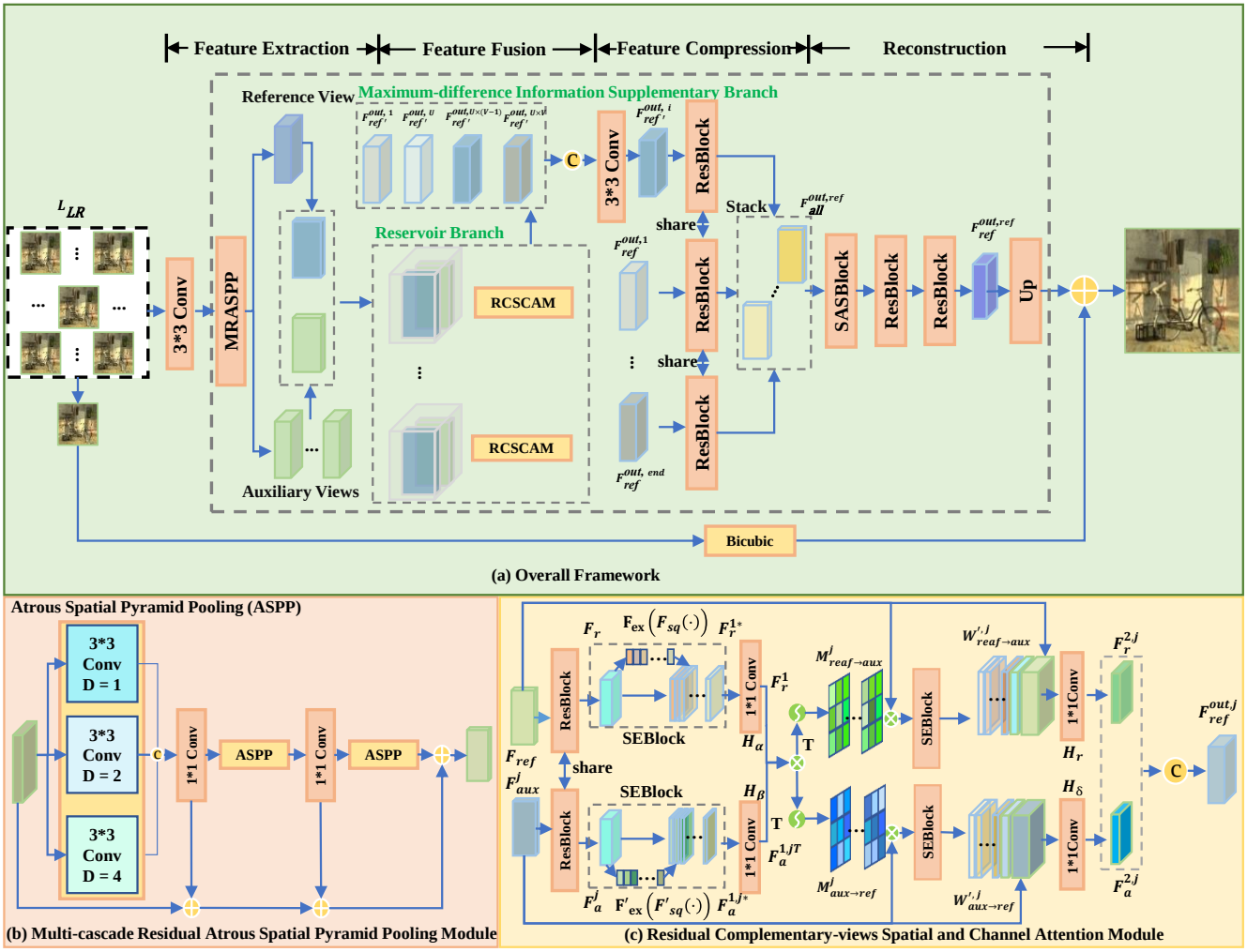


Figure 2. Network architecture of the proposed LF-CFANet. The overall network is composed of four parts (feature extraction, feature fusion, feature compression and reconstruction). The input of our network is SAIs, which is composed of a reference view image and auxiliary view images. The reference view is randomly selected from SAIs, and the remaining images are auxiliary view images. Finally, the output is a super-resolved reference view image. Note that, c is the concatenation operation.

3.1. Structure of 4D LF

The parameterization of a 4D LF usually consists of two parallel planes. These planes can accurately describe the light rays $\mathcal{L}(\Pi, \Omega)$. Each light ray intersects with the two planes, which are respectively called a spatial plane and an angular plane. As shown in Fig. 1(b), the spatial plane ($\Pi = (h, w)$) and angular plane ($\Omega = (u, v)$) are used to describe the structure of a 4D LF. Thus, one perspective of 4D LF images can be described by fixing Ω. Similarly, different views of one 3D scene can be described with a fixed Π.

By fixing (w^*, v^*) and varying (h, u), the epipolar-plane images (EPI) ($I_{w^*,v^*}(h, u)$) can be obtained. In the same way, $I_{h^*,u^*}(w, v)$ can be obtained. EPI can intuitively reflect the position changes of the objects in LF images from different views. The slope of EPIs represents the depth information of an object, so it can reflect the geometry structure of LF. At the same time, the integrity of the slope is an important evaluation criterion for judging whether the result

Datasets	Training	Test	LF Disparity	Scene
EPFL[38]	70	10	[-1,1]	Real-world
HCInew[39]	20	4	[-4,4]	Synthetic
HCIold[40]	10	2	[-3,3]	Synthetic
INRIA[41]	35	5	[-1,1]	Real-world
STFgantry[42]	9	2	[-7,7]	Real-world
STFlytro[43]	250	50	-	Real-world
Total	394	73		

Table 1. Public LF datasets used in our experiments

of LFSR maintains the geometric structure.

3.2. Feature extraction

The quality of discriminative features with rich context information is very useful to SR reconstruction. This information can be acquired by using a multi-scale receptive field and feature learning. Therefore, the feature-extraction module of our LF-CFANet follows [19], [44] and uses atrous spatial pyramid pooling (ASPP) module to extract the LF image features.

Fig. 2 shows the overall network architecture of the proposed LF-CFANet. As what can be seen, the input L_{LR} is composed of SAIs. The initial features (with 64 channels) of L_{LR} are extracted by a 3×3 convolution which is shown in Fig. 2(a), and then we use the multi-cascaded residual ASPP (MRASPP) module shown in Fig. 2(b) for multi-scale feature extraction to support the following parts. Specifically, the initial features of the LF views are first fed to the ASPP blocks, and the weights are shared for each view in these blocks. For each ASPP block, it is composed of three different dilated convolutions with a Leaky ReLU layer. These dilated convolutions, with dilation rates (D) 1, 2 and 4, are used to extract L_{LR} features with different receptive field. After a Leaky ReLU layer, we concatenate three output features and compress the number of channels through 1×1 convolution to make them more compact. These ASPP blocks not only obtain multi-receptive fields without changing the size of the feature maps, but also enrich the diversity of the convolutions. After three cascaded residual ASPP blocks, the extracted feature of each view is generated. These features can be specifically expressed as:

$$\{F^{i}_{\mathbf{each}} | i = 1, 2, ..., n\} = f_0\left(L_{\mathbf{LR}}\right), \tag{1}$$

where f_0 represents the MRASPPBlock and n is the number of SAIs.

For the output of MRASPPBlock ($F^{n}_{\mathbf{each}}$), the reference feature is randomly selected from the number of n output feature, and the auxiliary features are composed of the remaining features. These two types of features can be specifically expressed as:

$$\begin{aligned} F_{\mathbf{ref}} &= F^{i}_{\mathbf{each}} \\ F^{j}_{\mathbf{aux}} &= F^{j}_{\mathbf{each}}, \end{aligned} \tag{2}$$

where $i, j\ (1 \le i, j \le U \times V, i \ne j, i + j = n)$ represent the angular positions. The number of features i is one, while the number of features j is $U \times V - 1$.

As shown in Fig. 2(a), we concatenate $F_{\mathbf{ref}}$ and each $F^{j}_{\mathbf{aux}}$ to form a feature pair $\left\{F_{\mathbf{ref}}, F^{j}_{\mathbf{aux}}\right\}$. The way of selecting the complementary-view pairs makes our model more compatible for all views and increases the generalization performance of the network.

Algorithm 1 Squeeze and excitation blocks

Require:
The feature pair $\left\{F_{\mathbf{r}}, F^{j}_{\mathbf{a}}\right\} \in \mathbb{R}^{h \times w \times 64}$

1: **Squeeze:** The feature $(F_{\mathbf{r}}, F^{j}_{\mathbf{a}})$ compression is performed along the spatial dimension.
For each channel, compute
$\mathbf{F}^{1}_{sq}\left(F_{\mathbf{r}}\right) = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} F_{\mathbf{r}}(w, h)$;
For each channel, compute
$\mathbf{F}^{1,j}_{sq}\left(F^{j}_{\mathbf{a}}\right) = \frac{1}{W \times H} \sum_{i=w}^{W} \sum_{h=1}^{H} F^{i}_{\mathbf{a}}(w, h)$;
Each two-dimensional (H, W) feature channel becomes a number, which has a global receptive field.

2: **Excitation and Reweight:** Each feature channel generates a weight to represent the importance of the feature channel. The weight of the output of Excitation is regarded as the importance of each feature channel, and it is applied to each channel by multiplication.
Compute $F^{1\star}_{\mathbf{r}} = \mathbf{F}^{1}_{ex}\left(\mathbf{F}^{1}_{sq}\left(F_{\mathbf{r}}\right)\right)$;
Compute $F^{1,j\star}_{\mathbf{a}} = \mathbf{F}^{1,j}_{ex}\left(\mathbf{F}^{1,j}_{sq}\left(F^{j}_{\mathbf{a}}\right)\right)$;

Ensure:
$\left\{F^{1\star}_{\mathbf{r}}, F^{1,j\star}_{\mathbf{a}}\right\} \in \mathbb{R}^{h \times w \times 64}$

3.3. Residual complementary-view spatial and channel attention module (RCSCAM) in reservoir branch

The feature fusion part includes two branches. The first branch is a reservoir branch, and the second branch is MDISB. The reservoir branch is the key to fuse auxiliary-view information to reference-view information by using RCSCAM. Inspired by the stereo-attention mechanisms [20, 45] and spatial–temporal co-occurrence constraints [46–48], we develop an RCSCAM to supplement the sub-pixel information of the reference view.

As shown in Fig. 2(c), the input pair of features $\left\{F_{\mathbf{ref}}, F^{j}_{\mathbf{aux}}\right\}$ are separately fed to two ResBlocks (f_1) with 64 channels. These two ResBlocks share the same weight. The output features of f_1 are $F_{\mathbf{r}}, F^{j}_{\mathbf{a}} \in \mathbb{R}^{H \times W \times 64}$.

To explore the correlation among feature channels, we introduce SEBlocks following [18]. The pseudo-code to capture the channel attention is provided in Algorithm 1. This block processes the fed feature in three steps, the squeeze process, the excitation process, the reweight process. $F_{\mathbf{r}}$ and $F^{j}_{\mathbf{a}}$ are respectively fed to globally adaptive pooling $(\mathbf{F}^{1}_{sq}, \mathbf{F}^{1,j}_{sq})$ to obtain feature channels with 1×64 aggregated information. To capture channel-wise dependencies, two fully-connected (FC) layers are used. Then, the output weight of the excitation process represents the importance of the feature channel. They are applied to each channel by multiplication. These processes are denoted as $(\mathbf{F}^{1}_{ex}, \mathbf{F}^{1,j}_{ex})$. Then, the output are separately fed to 1×1 convolutions to generate the feature maps $(F^{1}_{\mathbf{r}}, F^{1,j}_{\mathbf{a}})$. These output can be specifically expressed as:

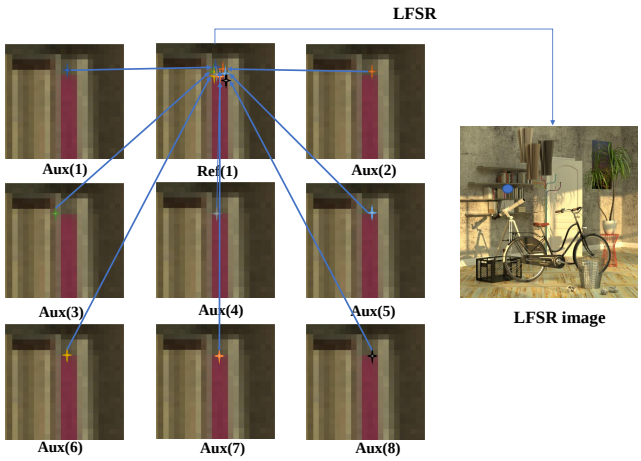


Figure 3. An illustration of supplement sub-pixel information. Here, a L_{LR} ($\mathbb{R}^{3\times3\times w\times h}$) is used as an example. We randomly choose a reference view ($U=2, V=1$), and remaining views are used as auxiliary views. For better comprehension, sub-pixel information from different auxiliary views is visually represented as stars with different colors. Note that, the information is added to the blue dot in LFSR image.

$$\begin{aligned} F_{\mathbf{r}}^{1} &= H_{\alpha}\left(f_{SE1}\left(F_{\mathbf{r}}\right)\right) \\ F_{\mathbf{a}}^{1,j} &= H_{\beta}\left(f_{SE2}\left(F_{\mathbf{a}}^{j}\right)\right), \end{aligned} \tag{3}$$

where f_{SE1}, f_{SE2} represent the SEBlocks, H_{α} and H_{β} respectively represent the 1×1 convolutions.

To generate a reference-auxiliary attention map, $F_{\mathbf{a}}^{1,j}$ is first transposed to $F_{\mathbf{a}}^{1,j\mathrm{T}}$, and then geometry-aware matrix is multiplied by matrix $F_{\mathbf{r}}^{1}$. The output of multiplying these two matrices is processed by the softmax to produce the final attention maps ($\mathcal{M}_{aux\rightarrow ref}^{j} \in \mathbb{R}^{H\times W\times W}$). Similarly, $\mathcal{M}_{ref\rightarrow aux}^{j}$ is generated. This process can be expressed as follows:

$$\begin{aligned} \mathcal{M}_{ref\rightarrow aux}^{j} &= \mathrm{S}\left(F_{\mathbf{r}}^{1} \otimes F_{\mathbf{a}}^{1,j\mathrm{T}}\right) \\ \mathcal{M}_{aux\rightarrow ref}^{j} &= \mathrm{S}\left(F_{\mathbf{r}}^{1} \otimes F_{\mathbf{a}}^{1,j\mathrm{T}}\right)^{T}, \end{aligned} \tag{4}$$

where $\otimes$ is the operation of batch-wise matrix multiplication, S and T denote the softmax and transposition, respectively.

To achieve information interaction of features between the reference-view and the auxiliary-view, $\mathcal{W}_{ref\rightarrow aux}^{j}$ and $\mathcal{W}_{aux\rightarrow ref}^{j}$ are generated by multiplying the input pair of features ($F_{\mathbf{ref}}, F_{\mathbf{aux}}^{j}$) and the attention maps ($\mathcal{M}_{ref\rightarrow aux}^{j}$, $\mathcal{M}_{aux\rightarrow ref}^{j}$), respectively. Both $\mathcal{W}_{ref\rightarrow aux}^{j}$ and $\mathcal{W}_{aux\rightarrow ref}^{j}$ contain the reference-view and auxiliary-view information. $\mathcal{W}_{ref\rightarrow aux}^{j}$ and $\mathcal{W}_{aux\rightarrow ref}^{j}$ can be calculated as:

$$\begin{aligned} \mathcal{W}_{ref\rightarrow aux}^{j} &= \mathcal{M}_{ref\rightarrow aux}^{j} \otimes F_{\mathbf{ref}} \\ \mathcal{W}_{aux\rightarrow ref}^{j} &= \mathcal{M}_{aux\rightarrow ref}^{j} \otimes F_{\mathbf{aux}}^{j}. \end{aligned} \tag{5}$$

As shown in Fig. 2, these two features ($\mathcal{W}_{ref\rightarrow aux}^{j}$, $\mathcal{W}_{aux\rightarrow ref}^{j}$) are fed into two new SEBlock to generate new features ($\mathcal{W}_{ref\rightarrow aux}^{\prime,j}$, $\mathcal{W}_{aux\rightarrow ref}^{\prime,j}$), respectively.

To retain the original features of reference and auxiliary view, the input pair of features ($F_{\mathbf{ref}}, F_{\mathbf{aux}}^{j}$) is concatenated with ($\mathcal{W}_{ref\rightarrow aux}^{\prime,j}, \mathcal{W}_{aux\rightarrow ref}^{\prime,j}$), respectively, and then fed to another 1×1 convolution. This process can be expressed as follows:

$$\begin{aligned} F_{\mathbf{r}}^{2,j} &= H_{\gamma}\left(cat\left(F_{\mathbf{ref}}, \mathcal{W}_{ref\rightarrow aux}^{\prime,j}\right)\right) \\ F_{\mathbf{a}}^{2,j} &= H_{\delta}\left(cat\left(F_{\mathbf{aux}}^{j}, \mathcal{W}_{aux\rightarrow ref}^{\prime,j}\right)\right), \end{aligned} \tag{6}$$

where cat is the concatenation operator, H_{γ} and H_{δ} represent the 1×1 convolutions to fuse these two types of features, respectively. $F_{\mathbf{r}}^{2,j}$ and $F_{\mathbf{a}}^{2,j}$ respectively represent the fully fused features of each pair by our RCSCAM.

The interacted features of complementary views are generated in this process. These four SEBlocks can continually distillate the valid information of the reconstruction. The result ($F_{\mathbf{ref}}^{out,j}$), fully integrating the complementary information, can be expressed as:

$$F_{\mathbf{ref}}^{out,j} = cat\left(F_{\mathbf{a}}^{2,j}, F_{\mathbf{r}}^{2,j}\right). \tag{7}$$

In the training process, the reference view feature ($F_{\mathbf{ref}}$) is generated by randomly selecting from initial features. Due to complex geometrical structure of LF images, the fusion features ($F_{\mathbf{ref}}^{out,j}$) obtained by RCSCAMs contain not only complementary information, but also local similarity information from different auxiliary views. The principle of RCSCAM is to obtain the feature similarities to all possible disparities between each pixel in the reference view and auxiliary view to generate an attention map. By introducing the attention mechanism, it makes the complementary information fully fused through feature-level information interaction for reconstructing SR. The effectiveness of RCSCAM is demonstrated in Section 4.3.

3.4. Maximum-difference information supplementary branch (MDISB)

As the second branch of feature fusion, the MDISB is used to select four maximum-difference fusion features for guiding the reference view reconstruction. This branch is to choose the four fusion features with the maximum-difference information relative to the reference view from the reservoir. After the RCSCAM, each pair of between reference view and auxiliary views generates one fusion feature. The number of fusion features is $n_{\mathbf{1}} = U\times V-1$ in total. Due to the parallax structure of the LF, the difference information of each auxiliary view is different to supplement reference-view information. The four angular-position initial features $[F_{\mathbf{each}}^{1}, F_{\mathbf{each}}^{U}, F_{\mathbf{each}}^{U\times(V-1)}, F_{\mathbf{each}}^{U\times V}]$ generated

Methods	Scale	EPFL	HCInew	HCIold	INRIA	STFgantry	STFlytro
Bicubic	×2	29.50/0.935	31.69/0.934	37.46/0.978	31.10/0.956	30.82/0.947	33.02/0.950
VDSR[49]	×2	32.01/0.959	34.37/0.956	40.34/0.985	33.80/0.972	35.80/0.980	35.91/0.970
EDSR[26]	×2	32.86/0.965	35.02/0961	41.11/0.988	34.61/0.977	37.08/0.985	36.84/0.975
GB[32]	×2	31.22/0.959	35.25/0969	40.21/0.988	32.76/0.972	35.44/0.983	35.04/0.956
RCAN[30]	×2	33.46/0.967	35.56/0.963	41.59/0.989	35.18/0.978	38.18/0.988	37.32/0.977
SAN[31]	×2	33.36/0.967	35.51/0.963	41.47/0.989	35.15/0.978	3798/0.987	37.26/0.976
LFBMD5D[33]	×2	31.15/0.955	33.72/0.955	39.62/0.985	32.85/0.969	33.55/0.972	35.01/0.966
resLF[16]	×2	33.22/0.969	35.79/0.969	42.30/0.991	34.86/0.979	36.28/0.985	35.80/0.970
LFSSR[17]	×2	34.15/0.973	36.98/0.974	43.29/0.993	35.76/0.982	37.67/0.989	37.57/0.978
LF-ATO[36]	×2	34.49/0.976	37.28/0.977	43.76/0.994	36.21/0.984	39.06/0.992	38.27/0.982
LF-InterNet[37]	×2	34.76/0.976	37.20/0.976	44.65/0.995	36.64/0.984	38.48/0.991	38.81/0.983
Ours	×2	34.92/0.976	37.46/0.977	44.16/0.994	36.81/0.985	39.48/0.992	38.91/0.983
Bicubic	×4	25.14/0.831	27.61/0.851	32.42/0.934	2682/0.886	25.93/0.843	27.84/0.855
VDSR[49]	×4	26.82/0.869	29.12/0.876	34.01/0.943	28.87/0.914	28.31/0.893	29.17/0.880
EDSR[26]	×4	27.82/0.892	29.94/0.893	35.53/0.957	29.86/0.931	29.43/0.921	30.29/0.903
GB[32]	×4	26.02/0.863	28.92/0.884	33.74/0.950	27.73/0.909	28.11/0.901	28.37/0.873
RCAN[30]	×4	28.31/0.899	30.25/0.896	35.89/0.959	30.36/0.936	30.25/0.934	30.66/0.909
SAN[31]	×4	28.30/0.899	30.25/0.898	35.88/0.960	30.29/0.936	30.25/0.934	30.66/0.909
LFBMD5D[33]	×4	26.61/0.869	29.13/0.882	34.23/0.951	28.49/0.914	28.30/0.900	29.07/0.881
resLF[16]	×4	27.86/0.899	30.37/0.907	36.12/0.966	29.72/0.936	29.64/0.927	28.94/0.891
LFSSR[17]	×4	29.16/0.915	30.88/0.913	36.90/0.970	31.03/0.944	30.14/0.937	31.21/0.919
LF-ATO[36]	×4	29.16/0.917	31.08/0.917	37.23/0.971	31.21/0.950	30.78/0.944	30.98/0.918
LF-InterNet[37]	×4	29.52/0.917	31.01/0.917	37.23/0.972	31.65/0.950	30.44/0.941	31.84/0.927
Ours	×4	29.58/0.917	31.24/0.918	37.24/0.972	31.89/0.951	31.05/0.948	31.99/0.928

Table 2. PSNR/SSIM values achieved by different methods for 2× and 4×SR, the best results are in red and the second best results are in blue

by the MRASPP block have maximum-difference information compared with the reference view. These four features concatenating with the reference-view feature are fed into the RCSCAM. The output of these four features through RCSCAM is $[F_{\mathbf{ref'}}^{out,1}, F_{\mathbf{ref'}}^{out,U}, F_{\mathbf{ref'}}^{out,U\times(V-1)}, F_{\mathbf{ref'}}^{out,U\times V}]$. Then, a concatenation operator (cat) is used to combine the output from RCSCAM. This MDISB process can be expressed as:

$$F_{\mathbf{ref'}}^{out,i} = cat\left(F_{\mathbf{ref'}}^{out,1}, F_{\mathbf{ref'}}^{out,U}, F_{\mathbf{ref'}}^{out,U\times(V-1)}, F_{\mathbf{ref'}}^{out,U\times V}\right), \tag{8}$$

where $F_{\mathbf{ref'}}^{out,i}$ represents the output of our MDISB for reference-view position, the input $[F_{\mathbf{ref'}}^{out,1}, F_{\mathbf{ref'}}^{out,U}, F_{\mathbf{ref'}}^{out,U\times(V-1)}, F_{\mathbf{ref'}}^{out,U\times V}]$ represent the fusion features that supplement the complementary-view information to $F_{\mathbf{ref}}$ by using our RCSCAM, respectively. As shown in Fig. 2(a), we concat these four features and compress them by a 3×3 convolution. The depth of this final feature is 64.

3.5. Feature compression

For feature compression, it can compress the feature depth to adapt to the part of reconstruction. We use ResBlocks to process each feature, which are $F_{\mathbf{ref}}^{out,1}, \ldots, F_{\mathbf{ref}}^{out,j}, F_{\mathbf{ref'}}^{out,i}, F_{\mathbf{ref}}^{out,j+1}, \ldots, F_{\mathbf{ref}}^{out,end}$ from two branches. All ResBlocks share the same parameters. We stack these features from all auxiliary views. These features are trained to integrate the complementary information from RCSCAM, and the maximum-difference information from MDISB. The output for feature compression can be written as:

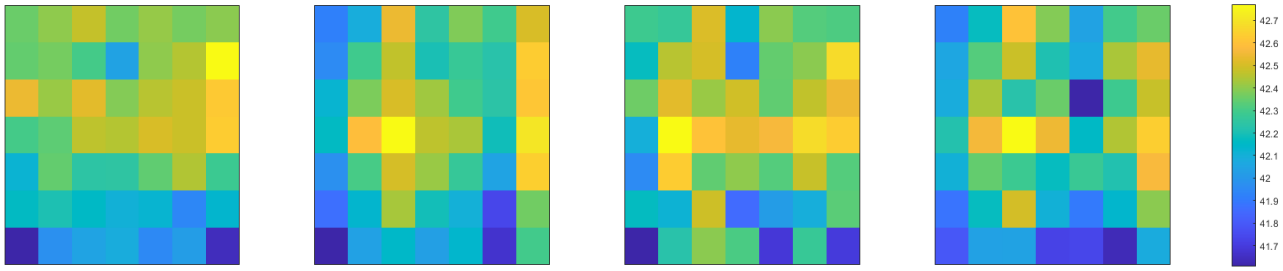
$$\begin{aligned} F_{\mathbf{all}}^{out,ref} = Stack[&F_{\mathbf{ref}}^{out,1}, \ldots, F_{\mathbf{ref}}^{out,j}, \\ &F_{\mathbf{ref'}}^{out,i}, \\ &F_{\mathbf{ref}}^{out,j+1}, \ldots, F_{\mathbf{ref}}^{out,end}], \end{aligned} \tag{9}$$

where $Stack$ is the operation of the feature stack.

3.6. Reconstruction

Inspired by the architecture of [49] in SISR, we set a similar structure to reconstruct the SR images. Following the method of [17], the feature $(F_{\mathbf{all}}^{out,ref})$ from the compression module firstly is reshaped and processed by a SASBlock. The process of SASBlock repeats 3 times to integrate angular and spatial domain information. The output feature is fed into two ResBlocks (with 64 channels). One ResBlock (two residual blocks) is for channel-wise view fusion. The other ResBlock (three residual blocks) is for

Figure 4. Visual comparisons of different methods on 2× / 4× reconstruction.



LFSSR
Max=40.12 Min=38.55
Avg=39.47 Std=0.313

LF-ATO
Max=42.60 Min=41.40
Avg=42.07 Std=0.267

LF-InterNet
Max=41.48 Min=39.84
Avg=40.84 Std=0.37

LF-CFANet
Max=42.77 Min=41.61
Avg=42.22 Std=0.27

Figure 5. Comparison of the PSNR of individual SAIs. Here, 7×7 input views are used to perform 2×LFSR. We use standard deviation (Std) to represent their uniformity.

channel fusion to generate the final reference-view feature ($F_{\mathbf{ref}}^{out,ref}$).

To reduce the memory consumption and computational complexity, we utilize an up-sampling Block $Up(\cdot)$ to increase the resolution of the reference-view image (L_{SR}^{ref}). This block, inspired by [16], is composed of a convolution layer, a shuffle layer and a convolution layer in order. Finally, L_{SR}^{ref} is generated by adding the residual map with the upsampled image. The reconstruction image from one angular position can be expressed as:

$$L_{\mathrm{SR}}^{ref} = Up\left(F_{\mathbf{ref}}^{out,ref}\right), \tag{10}$$

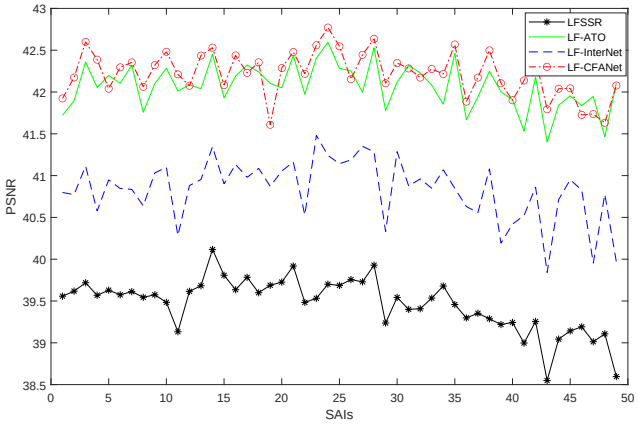


Figure 6. Comparison of the PSNR of individual SAIs.

where Up represents the process of reconstruction.

In order to be comprehensive, we use Fig. 3 to simplify our network on the LFSR process. Specifically, we first randomly select a view as the reference view, and then input all the views into our network. Through feature extraction, feature fusion, feature compression and reconstruction in our network, these processes can fully learn the differential subpixel information from the auxiliary views. The information can be added to the reference view for reconstruction.

4. Experiments

In this section, we first introduce the datasets (real-world scenes, synthetic scenes) and the implementation details. Then, we compare our LF-CFANet to several state-of-the-art SISR and LFSR methods. Finally, we conduct ablation studies to investigate the performance of our network with individual component modules.

4.1. Experimental Setup

4.1.1 Datasets

LF images are divided into two categories: synthetic datasets and real-world datasets. For real-world datasets, they are captured by different devices with different baseline lengths. Therefore, it is more meaningful for LF algorithms to adapt different datasets with different baseline lengths. As listed in Tab. 3.1, 6 public LF datasets (EPFL [38], HCInew [39], HCIold [40], INRIA [41], STFgantry [42] and STFlytro [43]) were used for training and testing in our experiments. There were a total of 394 LF scenes for training and 73 LF scenes for testing. For these datasets, the properties are different. Specifically, EPFL, INRIA and STFlytro are composed of rich outdoor-scene images captured with a Lytro Illum camera. HCInew, HCIold and STFgantry contain indoor LF images. The LF disparity of these datasets is multifarious and the angular resolution is 9×9 for each LF datasets. For both training and testing, low-resolution LF images were generated by the bicubic interpolation method.

4.1.2 Implementation Details

In our network, we have two types of convolutional layers, which are 3×3 and 1×1. All the 3×3 convolutional layers are zero-padded to keep the size of spatial resolution, and we set the number of Resblocks to 2,2,3 residual blocks in order. The feature depths of residual blocks are all 64.

In the training stage, we randomly crop the input LF images with the spatial size of 64×64. These cropped LF images are randomly processed by flipping the images horizontally or vertically, or rotating them 90 degrees. The factor of r is 2 or 4, and we respectively train the network with different factors. We train our network with Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The initial learning rate was set to $1e^{-4}$ and decreased by a factor of 0.5 every 250 epochs. In particular, the training of the full LF-CFANet is stopped after 600 epochs.

4.2. Comparison with State-of-the-Art Methods

We compared our LF-CFANet with recent state-of-the-art SISR and LFSR: Bicubic, VDSR [49], EDSR [26], GB [32], RCAN [30], SAN [31], LFBMD5D [33], resLF [16], LFSSR [17], LF-ATO [36] and LF-InterNet [37]. For a fair comparison these methods have been re-trained on the same training dataset as our method. Meanwhile, we chose bicubic interpolation as a baseline for comparison. For evaluation metrics, Peak Signal-to-Noise Ratio (PSNR) and structural similarity (SSIM) were used as quantitative metrics for performance evaluation. The higher value of these two metrics denotes the better LF reconstruction performance.

4.2.1 Quantitative Comparisons Results

The quantitative metrics (PSNR/SSIM) with 5×5 angular resolution for the 6 testing datasets are listed in Tab. 3.3. Note that, our method achieved higher PSNR and SSIM than SISR method RCAN[30]. Specifically, our method had an average increase of 1.7dB (×2) and 1.2dB (×4) higher in PSNR on the testing datasets. That is because, complementary information can be used effectively under the structure of LF. Moreover, our method achieved the best results in real-world datasets (EPFL, INRIA, STFgantry) and synthetic datesets (HCInew, HCIold). That is because, our LF-CFANet is based on the feature fusion driven by the attention mechanism, thus it is sensitive to disparity.

Due to different angular resolutions, the PSNR of each view in SAIs are not identical. As shown in Fig. 5 and Fig. 6, we also give the comparison of the PSNR of individual SAIs among LFSSR, ATO, LF-InterNet and our

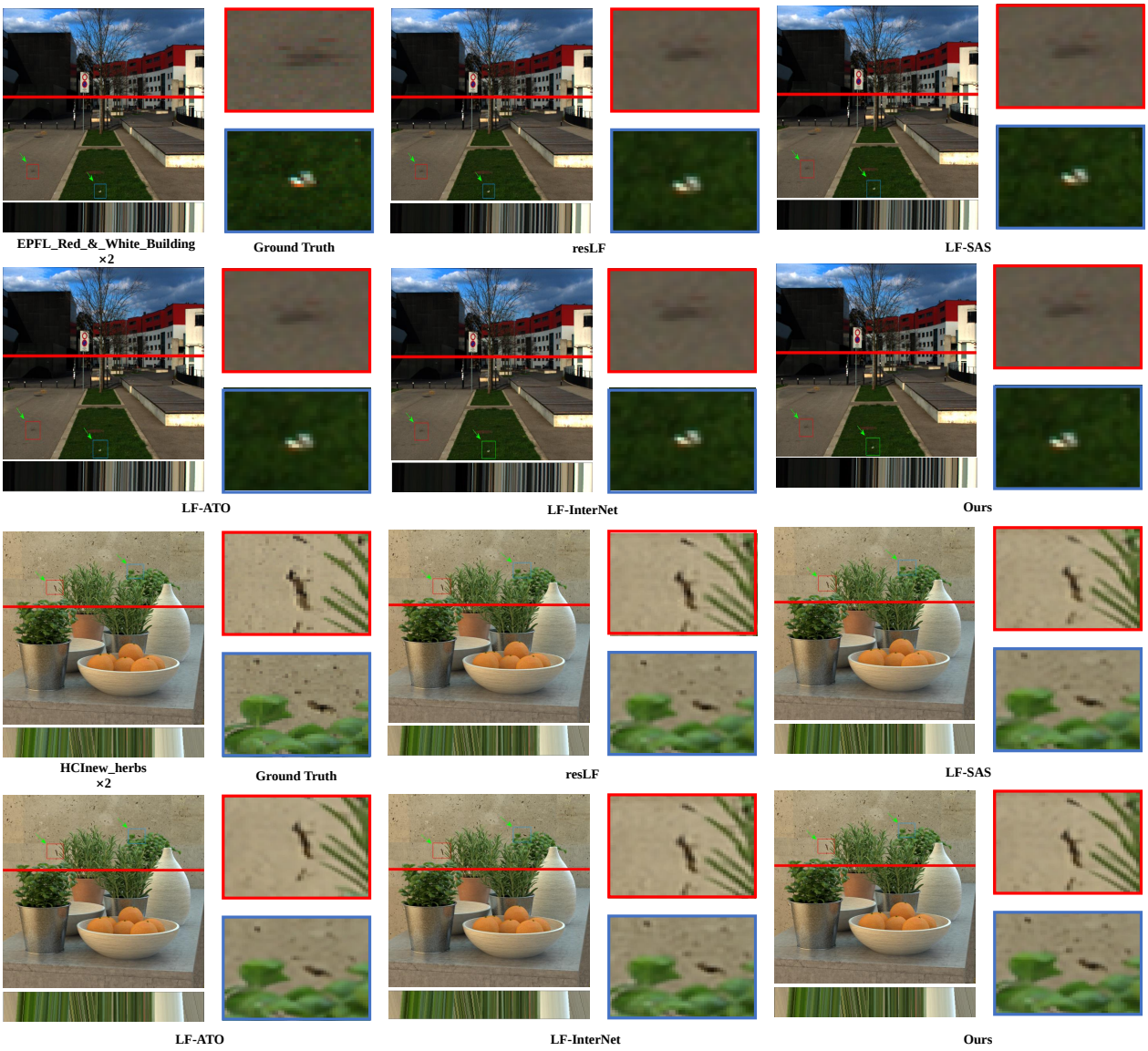


Figure 7. Visual comparisons of different methods on ×2 reconstruction for both synthetic and real-word scenes. The predicted central SAIs, the zoom-in of the framed patches, the EPIs at the colored lines. Zoom in the figure for better viewing.

method. Comparing with the same many-to-one approach (LF-ATO), our approach shows the significant performance improvements, as illustrated in Fig. 6. Although LFSSR, LF-ATO and LF-InterNet can use the angular information from all input views to super-resolve each view, our method can be observed that the gap among maximum-difference views of our method is much smaller than that of other methods. That is because, our method obtains the improvement by introducing MDISB to decrease the information degradation of maximum-difference views. The reconstruction quality of LF-CFANet is slightly higher than those of LFSR methods. Moreover, the computational efficiency of some state-of-the-art methods (LF-ATO, LF-InterNet) is demonstrated in Tab. 4.2.1. Note that, our method consumes little computational efficiency but achieves the best results.

4.2.2 Qualitative Comparisons Results

We provide the visual comparisons of different methods, as shown in Fig. 4 for ×2 and ×4 LFSR. Our LF-CFANet is able to restore the fine details and textures, such as the wheels in STFlytro_building_30. However, the comparison methods lost most high-frequency details in the reconstruction results. Comparing with our method, VDSR and SAN, state-of-the-art SISR methods, had poor details, because

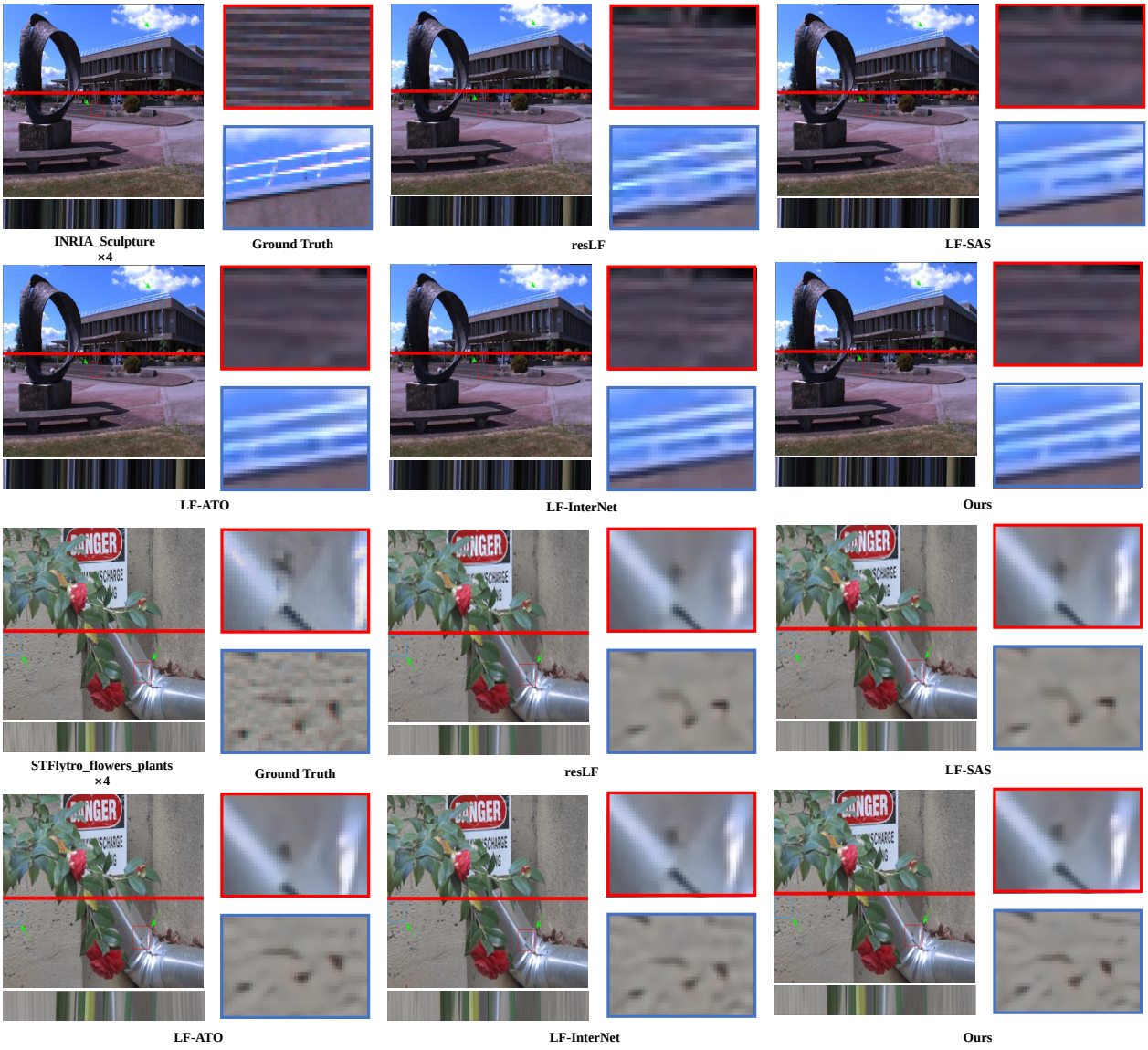


Figure 8. Visual comparisons of different methods on ×4 reconstruction for real-word scenes. The predicted central SAIs, the zoom-in of the framed patches, the EPIs at the colored lines. Zoom in the figure for better viewing.

both of them lack the complementary information to supplement image reconstruction. Although resLF, LF-SAS, LF-ATO and LF-InterNet methods could generate better results than SISR methods, they are not efficient to make use of complementary information in the process of LFSR. Our method can effectively and efficiently reconstruct LF images by introducing a channel and spatial attention mechanism. The more visual comparisons about LF parallax structure of LFSR methods are shown in Fig. 7 and 8. It can be seen that our method has clearer and more straight lines compared with the other LFSR methods. It was also be demonstrated that our method could preserve the structure characteristics of LF.

4.3. Ablation Study

In this subsection, we implement several experiments to investigate the effect of performance with different architectures.

4.3.1 Effectiveness of the MRASPP for feature extraction

The MRASPP is used to extract discriminative features. In order to make the experiment more convincing, we use LF-CFANet-onlyMRASPP and LF-CFANet-rmMRASPP to prove the effectiveness of the MRASPP. The results are

Architecture	PSNR	SSIM	Parameter
Bicubic	33.02	0.950	—
LF-CFANet-onlyMRASPP	38.62	0.980	2.32M
LF-CFANet-rmMRASPP	38.85	0.982	2.63M
LF-CFANet-onlyMDISB	38.71	0.982	2.54M
LF-CFANet-rmMDISB	38.87	0.983	2.41M
LF-CFANet-onlyRCSCAM	38.71	0.982	2.05M
LF-CFANet-rmRCSCAM	38.78	0.983	2.91M
LF-CFANet	**38.91**	**0.983**	**3.00M**

Table 3. The comparison results of different architectures of LF-CFANet on the dataset STFlytro with upscaling factor ×2. The result of bicubic method is listed as baseline. Note that, the meaning of onlyMRASPP, onlyMDISB and onlyRCSCAM is that only MRASPP, MDISB and RCSCAM block are used in our LF-CFANet, respectively. Meanwhile, the meaning of rmMRASPP, rmMDISB and rmRCSCAM is that only MRASPP, MDISB and RCSCAM block are removed in our LF-CFANet, respectively.

Network	Parameters(M)	FOPs(G)	Time(s)	PSNR(dB)
LF-ATO [36]	1.36	28.08	28.03	31.08
LF-InterNet [37]	4.80	47.46	52.21	31.01
LF-CFANet	3.00	51.14	63.04	31.24

Table 4. Comparisons of parameters, FLOPs and Times for ×4. FLOPs are calculated on 5×5×32×32. Time is calculated in an LF dataset

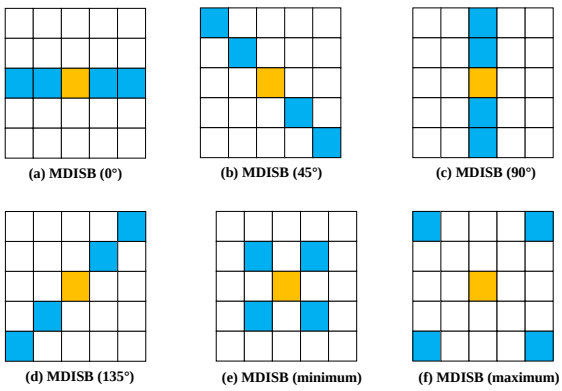


Figure 9. An illustration for supplement different information in MDISB. To be fair, we fix the angular position of reference view. The reference view is indicated in yellow. The blue blocks represents the angular positions of auxiliary views, which are used to supplement different information in MDISB.

shown in Tab. 4.2.1. As expected, LF-CFANet-rmMRASPP suffered a 0.06 dB decrease after removing MRASPP. That is because, our MRASPP extracted features with different scales, which can make the feature representations more robust. Meanwhile, discriminative features with rich context information can be extracted by using multiple receptive field of atrous convolutions. Therefore, our model can obtain accurate features to reconstruct LF.

Architecture	PSNR	SSIM
MDISB (0°)	38.83	**0.983**
MDISB (45°)	38.84	**0.983**
MDISB (90°)	38.81	**0.983**
MDISB (135°)	38.89	**0.983**
MDISB (minimum)	38.87	0.964
MDISB (maximum)	**38.91**	**0.983**

Table 5. THE COMPARISON RESULTS OF DIFFERENT SUPPLEMENT INFORMATION OF MDISB ON THE DATASET STFLYTRO WITH UPSCALING FACTOR ×2.

4.3.2 Effectiveness of the MDISB for feature fusion

MDISB is used to guide the reference view reconstruction. To validate the effectiveness of the MDISB, we remove this block, and we show the results in Tab. 4.2.1. LF-CFANet-rmMDISB suffered a 0.04 dB decrease comparing with LF-CFANet. That is because, this block can strengthen the influence of angular-position features on the process of reconstructing the reference-view image. Recall that in Eq. (8), we select four angular-position features with maximum-difference information according to the structure of LF to enhance the influence of maximum-difference views.

As shown in Tab. 4.3.3 and Fig. 4.2.2, we also investigated the performance of MDISB with different angular positions for auxiliary views. The reconstruction accuracy consistently improved, as the degree of differentiated information increased. Tab. 4.3.3 shows that MDISB (maximum) had the best result. That is because, MDISB (0°), MDISB (45°), MDISB (90°) and MDISB (135°) just provide differentiated information in the same direction, while the difference information for MDISB (minimum) and MDISB (maximum) have four directions. Meanwhile, the four views in MDISB (maximum) are the furthest away from the angular position of reference view, and the maximum degree of differentiation can be provided for reconstructing the reference view.

4.3.3 Effectiveness of the RCSCAM for feature fusion

The RCSCAM plays a key role in our LF-CFANet. This model can enhance the complementary information exploitation capability between the reference view and complementary view by introducing the attention mechanism. For a comparative experiment, we just use feature concatenation to replace our RCSCAM. As shown in Tab. 4.3.3, the block had a significant influence on the result, and the PSNR suffered a 0.13 dB decrease. Without a spatial and channel attention mechanism, the complementary information from the cross-parallax image cannot be effectively learned to supplement the reference view.

5. Conclusions

In this paper, we address the LFSR problem by proposing the complementary-view feature attention network (LF-CFANet). The LF-CFANet is mainly to improve the fusion of complementary-view information, by using RCSCAM and MDISB. For RCSCAM, we use spatial and channel attention to effectively extract the complementary-view feature information to supplement the reference view. To guide the reference view reconstruction, MDISB is proposed to supplement the most differentiated feature-level information. As demonstrated in the experiments, MDISB works well in the process of reconstruction. In this way, the reference view image can be effectively and efficiently reconstructed. The experimental results demonstrate that our method achieves the state-of-the-art quantitative and qualitative SR performance in LF, and it is more robust to real-world scenes.

It is worth noting that the quality of the supplementary information from MDISB is crucial and improves the reconstruction accuracy. Therefore, a further study of the maximum-difference views is needed, and could possibly use less views to reconstruct the whole LF views. For future work, we will use the framework of encoder and decoder to improve the quality of feature fusion with fewer LF views. In this case, it can take a further step toward consumer applications.

References

[1] R. Ng. Lytro redefines photography with light field cameras, 2018. 1

[2] C. Perwa and L. Wietzke. Raytrix: Light filed technology, 2018. 1

[3] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.*, 32(4):73–1, 2013. 1

[4] H. Zhu, Q. Wang, and J. Yu. Occlusion-model guided antiocclusion depth estimation in light field. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):965–978, 2017. 1

[5] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu. Saliency detection via depth-induced cellular automata on light field. *IEEE Transactions on Image Processing*, 29:1879–1889, 2019. 1

[6] M. Zhang, J. Li, J. Wei, Y. Piao, H. Lu, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, and E. Fox. Memory-oriented decoder for light field salient object detection. In *NeurIPS*, pages 896–906, 2019. 1

[7] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu. Plenopatch: Patch-based plenoptic image manipulation. *IEEE transactions on visualization and computer graphics*, 23(5):1561–1573, 2016. 1

[8] T. E. Bishop and P. Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):972–986, 2011. 1, 3

[9] S. Wanner and B. Goldluecke. Spatial and angular variational super-resolution of 4d light fields. In *European Conference on Computer Vision*, pages 608–621. Springer, 2012. 1, 3

[10] Wanner, Sven and Goldluecke, Bastian. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):606–619, 2013. 1, 3

[11] K. Mitra and A. Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 22–28. IEEE, 2012. 1, 3

[12] Y. Yuan, Z. Cao, and L. Su. Light-field image superresolution using a combined deep cnn based on epi. *IEEE Signal Processing Letters*, 25(9):1359–1363, 2018. 1, 3

[13] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 24–32, 2015. 3

[14] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24(6):848–852, 2017. 3

[15] D. Li, D. Yang, S. Wang, and H. Sheng. Light field super-resolution based on spatial and angular attention. In *International Conference on Wireless Algorithms, Systems, and Applications*, pages 314–325. Springer, 2021. 1

[16] S. Zhang, Y. Lin, and H. Sheng. Residual networks for light field image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11046–11055, 2019. 1, 3, 7, 8, 9

[17] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018. 1, 7, 9

[18] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2, 5

[19] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019. 5

[20] Y. Wang, X. Ying, L. Wang, J. Yang, W. An, and Y. Guo. Symmetric parallax attention for stereo image super-resolution. *arXiv preprint arXiv:2011.03802*, 2020. 2, 5

[21] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019. 2

[22] Z. Wang, J. Chen, and S. C. Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[23] Dong, Chao and Loy, Chen Change and He, Kaiming and Tang, Xiaoou. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and*

machine intelligence, 38(2):295–307, 2015. 2, 3

[24] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2, 3

[25] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[26] B. Lim, S. Son, H. Kim, S. Nah, and 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*, pages 136–144, 2017. 3, 7, 9

[27] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. 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*, pages 114–125, 2017. 3

[28] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 3

[29] Zhang, Yulun and Tian, Yapeng and Kong, Yu and Zhong, Bineng and Fu, Yun. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. 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. 3, 7, 9

[31] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 3, 7, 9

[32] M. Rossi and P. Frossard. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Transactions on Image Processing*, 27(9):4207–4218, 2018. 3, 7, 9

[33] M. Alain and A. Smolic. Light field super-resolution via lfbm5d sparse coding. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 2501–2505. IEEE, 2018. 3, 7, 9

[34] Y. Huang, W. Wang, and L. Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 235–243, 2015. 3

[35] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan. Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Transactions on Image Processing*, 27(9):4274–4286, 2018. 3

[36] J. Jin, J. Hou, J. Chen, and 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*, pages 2260–2269, 2020. 3, 7, 9, 12

[37] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo. Spatial-angular interaction for light field image super-resolution. In *European Conference on Computer Vision*, pages 290–308. Springer, 2020. 3, 7, 9, 12

[38] M. Rerabek and T. Ebrahimi. New light field image dataset. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, number CONF, 2016. 4, 9

[39] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2016. 4, 9

[40] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, volume 13, pages 225–226. Citeseer, 2013. 4, 9

[41] M. Le Pendu, X. Jiang, and C. Guillemot. Light field inpainting propagation via low rank matrix completion. *IEEE Transactions on Image Processing*, 27(4):1981–1993, 2018. 4, 9

[42] V. Vaish and A. Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7), 2008. 4, 9

[43] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein. Stanford lytro light field archive, 2016. 4, 9

[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5

[45] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020. 5

[46] H. Sheng, S. Wang, Y. Zhang, D. Yu, X. Cheng, W. Lyu, and Z. Xiong. Near-online tracking with co-occurrence constraints in blockchain-based edge computing. *IEEE Internet of Things Journal*, 8(4):2193–2207, 2020. 5

[47] H. Sheng, Y. Zhang, Y. Wu, S. Wang, W. Lyu, W. Ke, and Z. Xiong. Hypothesis testing based tracking with spatio-temporal joint interaction modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2971–2983, 2020.

[48] S. Wang, H. Sheng, Y. Zhang, Y. Wu, and Z. Xiong. A general recurrent tracking framework without real data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13219–13228, 2021. 5

[49] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 7, 9