Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

MPIN: a macro-pixel integration network for light field super-resolution*

Xinya WANG[1], Jiayi MA‡[1], Wenjing GAO[1], Junjun JIANG[2]

[1]*Electronic Information School, Wuhan University, Wuhan 430072, China*

[2]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*

E-mail: wangxinya@whu.edu.cn; jyma2010@gmail.com; wenjinggao@whu.edu.cn; junjun0595@163.com

Received Oct. 20, 2020; Revision accepted Mar. 26, 2021; Crosschecked Aug. 31, 2021

Abstract: Most existing light field (LF) super-resolution (SR) methods either fail to fully use angular information or have an unbalanced performance distribution because they use parts of views. To address these issues, we propose a novel integration network based on macro-pixel representation for the LF SR task, named MPIN. Restoring the entire LF image simultaneously, we couple the spatial and angular information by rearranging the four-dimensional LF image into a two-dimensional macro-pixel image. Then, two special convolutions are deployed to extract spatial and angular information, separately. To fully exploit spatial-angular correlations, the integration resblock is designed to merge the two kinds of information for mutual guidance, allowing our method to be angular-coherent. Under the macro-pixel representation, an angular shuffle layer is tailored to improve the spatial resolution of the macro-pixel image, which can effectively avoid aliasing. Extensive experiments on both synthetic and real-world LF datasets demonstrate that our method can achieve better performance than the state-of-the-art methods qualitatively and quantitatively. Moreover, the proposed method has an advantage in preserving the inherent epipolar structures of LF images with a balanced distribution of performance.

Key words: Light field; Super-resolution; Macro-pixel representation
https://doi.org/10.1631/FITEE.2000566 **CLC number:** TP312

1 Introduction

With the advent of the commercial portable light field (LF) camera, e.g., Lytro (https://www.lytro.com/) and Raytrix (https://www.raytrix.de/), LF imaging has facilitated a variety of applications, such as depth-sensing (Peng et al., 2018), post-capture refocusing (Ng et al., 2005), segmentation (Yücer et al., 2016), and material classification (Wang TC et al., 2016). By inserting the micro-lens array between the main lens and the imaging plane, LF imaging not only records the accumulated intensity at each image position (i.e., spatial information), but also separates intensity values for each ray direction (i.e., angular information) (Zhu et al., 2017; Yeung et al., 2019). However, offering rich view descriptions of the scene, LF images have relatively low spatial resolution due to the trade-off between spatial and angular resolution, which limits the range of potential development. In this paper, we focus on the task of LF spatial super-resolution (SR).

Because LF records multiple views in sub-aperture images (SAIs) with sub-pixel offsets, some traditional methods first adopt depth or disparity estimation techniques to warp or register SAIs and regularize the SR reconstruction process with different priors. However, these methods rely heavily on disparity estimation, any defect of which may result in significant artifacts. Meanwhile, the shallow

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61773295)

ORCID: Xinya WANG, https://orcid.org/0000-0003-2144-9811; Jiayi MA, https://orcid.org/0000-0003-3264-3265

heuristic model has limited expressive power, and fails to recover the complex image details. Recently, learning-based methods have emerged for LF SR. With the success of deep learning, several deep neural network models have demonstrated promising results for LF SR, in which different view combinations are employed to implicitly excavate corresponding relations. Since in these methods, parts of views are used to reconstruct SAIs of an LF image separately or in a certain sequence, they fail in fully exploiting both angular and spatial information. Consequently, their performances vary greatly in different SAIs.

To fully exploit both spatial and angular information, we merge these two kinds of information into the two-dimensional (2D) space to solve the LF SR problem in the macro-pixel representation, which is implemented by an integration network with angular shuffle, termed MPIN. As shown in Fig. 1, by collecting pixels from the same position of SAIs, we can obtain a position-specific macro-pixel representation, which is regarded as a collection to rearrange the LF image according to their coordinates. In this way, the four-dimensional (4D) LF image is reshaped into a 2D macro-pixel image (MPI), which not only couples spatial and angular information in the 2D space, but also alleviates the processing difficulty created by the high dimension of LF images. Taking 2D MPIs as input, in our method, we deploy two special convolutions with subtle parameter settings to extract spatial and angular information separately. To explore spatial-angular correlations, the integration resblock is designed to aggregate these two kinds of features for mutual guidance, by which spatial-angular coherence is better preserved. As features are extracted in a macro-pixel pattern, a novel angular shuffle layer is tailored to improve the spatial resolution of macro-pixel features without confusion. In particular, periodic shuffling is based on the macro-pixel of each spatial position as a unit to upscale the low-resolution (LR) features into a high-resolution (HR) space. Compared to other learning-based methods, we improve the spatial resolution in the macro-pixel domain, achieving a more balanced performance. By exploiting spatial-angular correlations with mutual guidance, our method can be angularly coherent with accurate spatial improvement.

The main contributions of this paper are as follows:

1. A novel and efficient network is proposed to address the LF spatial SR problem based on MPI. In particular, we couple the spatial and angular information into the 2D macro-pixel pattern for efficient processing. Since we super-resolve the whole LF image simultaneously, our method is superior to the state-of-the-art methods and has a balanced performance distribution.

2. We propose an integration resblock to mutually exploit spatial-angular correlations. To enhance the spatial resolution of MPI representation, an angular shuffle layer is designed for macro-pixel features, which can effectively avoid aliasing. In this way, inherent epipolar structures can be better preserved.

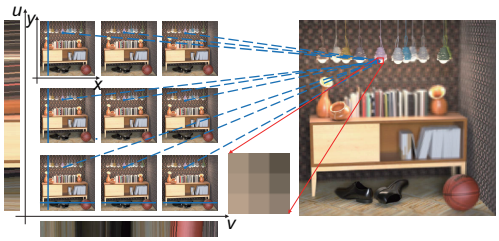


Fig. 1 Transformation from an array of sub-aperture images (SAIs) (left) to a macro-pixel image (MPI) (right). The pixels from the same position of SAIs are collected as a unit to reorganize the MPI. Epipolar plane images are extracted horizontally (bottom of SAIs) and vertically (left of SAIs) along the blue lines for visualization. References to color refer to the online version of this figure

2 Related work

Since LF images capture scenes with multiple views, most LF SR methods take advantage of the complementary information and the correlations between them, which can greatly improve the SR performance. Existing methods can be divided into three categories, projection-, optimization-, and learning-based.

Projection-based methods focus on projection and resample of LF data with the imaging principles of light field cameras. Lim J et al. (2009) first mapped the sub-pixel shift of redundant views from the angular space into the spatial space by projecting them onto convex sets. Focusing on the focal stack transformation problem, Nava and Luke (2009) exploited the refocusing principle and projected pixels from other views to the central view to obtain an all-in-focus image in a high-resolution space. Similarly, Pérez et al. (2012, 2015) proposed the Fourier

slice super-resolution to obtain the super-resolved discrete focal stack transform. In addition, Liang and Ramamoorthi (2015) demonstrated that typical lenslet LF cameras preserve frequency components above the spatial Nyquist rate and perform spatial super-resolution with the guidance of depth information to project the LF samples to the target view. To relieve the dependency of camera parameters and depth information, Wang YL et al. (2016) redefined the mapping function between the disparity of certain pixels and their shearing shift in projection-based methods. These projection-based methods consider only the internal similarity among different views and fail to restore rich high-frequency details.

Optimization-based methods use depth or disparity information, and the super-resolved LF images are reconstructed using various optimization frameworks based on different prior hypotheses. In Bishop et al. (2009) and Bishop and Favaro (2012), the depth map was explicitly recovered and the spatial LF SR problem was solved using a variational Bayesian framework with Lambertian reflectance priors (Li et al., 2020). In Mitra and Veeraraghavan (2012), a disparity-dependent Gaussian mixture model was proposed as an alternative, and the super-resolved LF images were reconstructed by a linear minimum mean squared error estimator. Moreover, Wanner and Goldluecke (2014) estimated the disparity maps from epipolar images with the structure tensor based method and conducted both spatial and angular SR in a variational optimization procedure. Rossi and Frossard (2017, 2018) coupled the multi-frame SR method with a graph regularizer to enforce the geometrical consistency of the LF image, which avoids the explicit disparity estimation. Inspired by LFBM5D used for light field denoising, a new method proposed by Alain and Smolic (2018) iteratively alternates between LFBM5D filtering and back-projection for LF SR. The performance of optimization-based methods is, to some extent, determined by the accuracy of depth information. Furthermore, the shallow heuristic model has a restricted capacity to reconstruct the complex structure.

Learning-based methods have been developed recently, and achieved remarkable performance. Farrugia et al. (2017) learned a linear mapping between the LR and HR in a low-dimensional subspace with ridge regression (RR). With the success of deep learning, LFCNN was the first convolutional neural network (CNN) based LF SR method introduced by Yoon et al. (2015), where they sent four tuples of SAI stacks into the SRCNN architecture (Dong et al., 2014) to jointly increase the spatial and angular resolution. To obtain better results, Fan et al. (2017) developed a two-stage CNN framework, in which different SAIs were aligned by patch matching in the first stage and a multi-patch fusion CNN was used in the second stage. Subsequently, a shallow CNN model was used to super-resolve the LF raw data directly from plenoptic cameras without decoding to SAIs in Gul and Gunturk (2018). Regarding an LF image as a sequence of 2D images, LFNet (Wang YL et al., 2018) was developed to model the spatial correlation between adjacent views in a bidirectional recurrent way and accumulate contextual information from multiple scales with a specially designed fusion layer. Inspired by epipolar geometry used for depth estimation, Zhang et al. (2019) grouped different image stacks into multiple branches to super-resolve the central view by residual learning in several position-specific models. Considering all views for LF SR, LFSAS was proposed by Yeung et al. (2019) to alternately shuffle LF features for alternate spatial-angular convolution. Jin et al. (2020) designed an all-to-one LF SR framework and performed structural consistency regularization to preserve the parallax structure among reconstructed views. In addition, Wang YQ et al. (2020) developed a spatial-angular interactive network where spatial and angular features were separately extracted and then interacted for progressive incorporation. Different from Wang YQ et al. (2020), in our method, the angular and spatial features are merged in the integration resblock to generate the most spatial-angular coherent features. Our method guides both spatial and angular residual extraction, whereas these two features were fused separately in Wang YQ et al. (2020). Because we perform super-resolution in the MPI representation, there is no need for the reshaping operation that was necessary in Wang YQ et al. (2020), which could be more efficient.

3 Problem formulation

Adopting the simplified 4D representation in Levoy and Hanrahan (1996), we denote the LF

image as $\boldsymbol{L} \in \mathbb{R}^{U\times V\times X\times Y\times C}$, where the pairs (U, V) and (X, Y) represent the angular and spatial dimensions respectively, and C denotes the number of RGB channels of the LF image. In general, an LF image can be regarded as an array of viewpoints of the scene with varying angular coordinates (u, v), which is called an SAI, $\boldsymbol{L}_{u,v} \in \mathbb{R}^{X\times Y\times C}$. To improve the spatial resolution of the whole image, some existing methods (Wang YL et al., 2018; Zhang et al., 2019) super-resolve each SAI by taking different view combinations as reference. As illustrated in Fig. 1, fixing one of the spatial coordinates and one of the angular coordinates, spatial-angular slices are obtained from the LF image and defined as epipolar plane images (EPIs) (e.g., $\boldsymbol{L}_{u,x} \in \mathbb{R}^{V\times Y\times C}$). Since the oblique lines are related to the disparity information among multiple views, the existing method (Yuan et al., 2018) also designs the post-processing networks for EPIs to compensate for the spatial-angular correlations.

Instead of focusing only on several dimensions of LF images, we restore the whole image by arranging the 4D LF image into a 2D MPI. Specifically, at the spatial position (x, y), the macro-pixel can be denoted as $\boldsymbol{L}_{x,y} \in \mathbb{R}^{U\times V\times C}$, which collects the angular information of this position. Usually, the angular resolution of the LF image is far smaller than its spatial resolution. Therefore, we place each macro-pixel at a specific spatial location according to its coordinates to reorganize the 4D LF image into a 2D MPI rather than an array of SAIs. Supposing that the angular dimension satisfies $U = V = A$, the original LF image $\boldsymbol{L} \in \mathbb{R}^{U\times V\times X\times Y\times C}$ can be folded as $\boldsymbol{M} \in \mathbb{R}^{AH\times AW\times C}$. Consequently, we solve the spatial SR problem in the macro-pixel domain.

Given the LR MPI $\boldsymbol{M}^{\mathrm{LR}} \in \mathbb{R}^{Ah\times Aw\times C}$ with angular resolution $A \times A$, where h and w denote the width and height of each view respectively, the goal of our method is to estimate the super-resolved MPI $\boldsymbol{M}^{\mathrm{SR}} \in \mathbb{R}^{AH\times AW\times C}$ with the enhancement of spatial resolution at the upscale factor s, in which $H = s \cdot h$ and $W = s \cdot w$. The formulation is

$$\boldsymbol{M}^{\mathrm{SR}} = \mathcal{F}(\boldsymbol{M}^{\mathrm{LR}}; \theta), \tag{1}$$

in which $\mathcal{F}(\cdot)$ represents the mapping function of the super-resolution model with parameter θ. Note that our proposed method always operates in the macro-pixel domain without the reshaping operation that was involved in Yeung et al. (2019) and Wang YQ et al. (2020).

4 Proposed method

We first introduce the design of our whole network. As illustrated in Fig. 2, our network consists of two parts, spatial-angular feature extraction (used to abstract both spatial and angular information) and feature fusion and reconstruction (used to merge the information and enlarge it).

4.1 Spatial-angular feature extraction network

In this network, we elaborately design the basic feature extraction modules because the MPI couples the spatial-angular information in a 2D plane. As illustrated in Fig. 2a, we first adopt parallel spatial convolution and angular convolution to process the LR MPI. Then, a series of integration resblocks are added to the network in a residual way, and they are supposed to fully extract both spatial correlations of the scene and inter-view angular correlations among rearranged LR SAIs.

1. Spatial convolution

To capture the spatial information from input MPI, we use the dilation convolution with a stride of one, where the dilation rate is equal to the angular resolution (A) of the original LF image. In addition, zero padding is used to ensure that the output has the same size as the input. In this way, spatial convolutions are expected to extract spatial correlations of scenes that are irrelevant to the angular information.

2. Angular convolution

When we arrange the LF image into the MPI representation, the pixels in one collection are observed at the same spatial position from all view directions. Thereby, we use the kernel size of $A \times A$ in the convolution to acquire the angular correlations among SAIs. To avoid aliasing, the stride of this convolution is set as A, and thus the output size is downsampled to $1/A$ of the input.

3. Integration resblock

Since the two basic convolutions are concentrated on their information separately without interaction, integration modules are needed to incorporate these spatial and angular features for mutual guidance. Therefore, inspired by Yi et al. (2019), we tailor an integration resblock to fully extract two

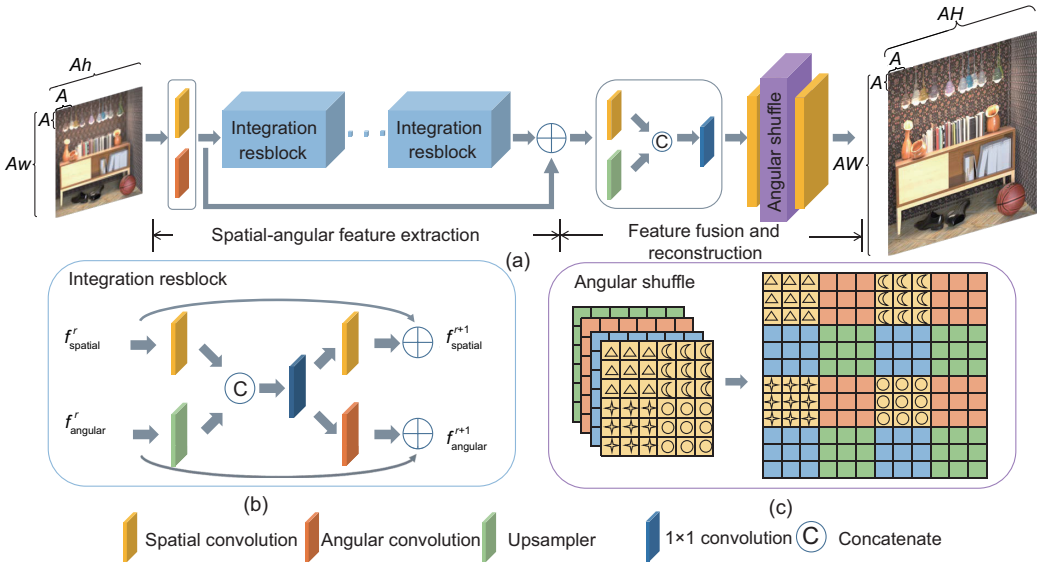


Fig. 2 Illustration of the proposed method: (a) the whole network of the proposed method with input size $Ah \times Aw$ and output size $AH \times AW$, where $H = s \cdot h$, $W = s \cdot w$ at upscale factor s; (b) the integration resblock for extracting spatial-angular features progressively; (c) a toy example for the angular shuffle operation with spatial resolution (of SAIs) of 2×2, angular resolution of 3×3, and upscale factor of 2

kinds of information. As shown in Fig. 2b, taking the r^{th} integration resblock with the total number of R as an example, a pair of spatial and angular features with N maps are fed into this block. First, as the spatial resolution of the angular feature is $1/A$ times that of the spatial feature, we deploy a spatial convolution and an upsampler to further acquire rather self-independent features (e.g., $\boldsymbol{I}^1_{\text{spatial}}$ and $\boldsymbol{I}^1_{\text{angular}}$) with the same channel number N, which can be described as

$$\begin{cases} \boldsymbol{I}^1_{\text{spatial}} = \mathcal{S}(\boldsymbol{f}^r_{\text{spatial}}), \\ \boldsymbol{I}^1_{\text{angular}} = \mathcal{U}(\boldsymbol{f}^r_{\text{angular}}), \end{cases} \tag{2}$$

in which $\mathcal{S}$ denotes the spatial convolution and $\mathcal{U}$ represents the upsampler consisting of a 1×1 convolutional layer and shuffle pixel operation (Shi et al., 2016) to restore the spatial size of the angular feature. Later, two kinds of features are concatenated and merged into one part, and thus this aggregated deep feature map contains both spatial and angular information. Naturally, we use one convolutional layer to distill this deep feature into a more concise one, which is supposed to be the most spatial-angular-coherent. Because feature maps are organized as MPI representation similarly, we set the kernel size as 1×1 to avoid distortion:

$$\boldsymbol{I}^2 = \mathcal{H}_{1\times1}([\boldsymbol{I}^1_{\text{spatial}}, \boldsymbol{I}^1_{\text{angular}}]), \tag{3}$$

where $\boldsymbol{I}^2$ is the distilled feature map with N channels, and $\mathcal{H}_{1\times1}$ represents the 1×1 convolution. After integration, we use the spatial-angular dependent features to guide both spatial and angular residual extraction. Thereby, we can obtain a pair of outputs from

$$\boldsymbol{f}^{r+1}_{\text{spatial}} = \mathcal{S}(\boldsymbol{I}^2) + \boldsymbol{f}^r_{\text{spatial}}, \tag{4}$$

$$\boldsymbol{f}^{r+1}_{\text{angular}} = \mathcal{A}(\boldsymbol{I}^2) + \boldsymbol{f}^r_{\text{angular}}, \tag{5}$$

in which $\mathcal{A}$ represents the angular convolution. Due to the contribution of mutual guidance, the updated spatial information is angular-correlated and the guided angular features are better preserved.

4.2 Feature fusion and reconstruction network

In this part, we first merge a pair of outputs from the feature extraction network and improve the spatial resolution to reconstruct the super-resolved MPI. Specifically, after one spatial convolution for spatial features and an upsampler for angular features, a 1×1 convolutional layer is used to fuse two kinds of features. Later, we deploy a spatial convolution and an angular shuffle layer, especially for MPI magnification. Finally, a spatial convolution is used for reconstruction. The angular shuffle layer is

detailed below.

Since the MPI collects pixels from different perspectives into an $A\times A$ spatial area, sharing the same insight of sub-pixel convolution (Shi et al., 2016), we design an efficient angular shuffle layer to improve the spatial resolution of feature maps in an MPI representation. As illustrated in Fig. 2c, angular shuffle regards the macro-pixel of each spatial position as a unit to shuffle pixels of MPI features, which means that the $A \times A$ pixels are rearranged as a collection and shuffled from channel dimension to spatial dimension periodically. In other words, the angular shuffle is the shuffle pixel based on the macro-pixel. Mathematically, this operation can be described as follows:

$$\mathcal{AS}(\boldsymbol{T})_{x,y,c} = \boldsymbol{T}_{n,m,p}, \tag{6}$$

where

$$\begin{aligned} x &= \lfloor n/(As) \rfloor \cdot A + \mathrm{mod}(\mathrm{mod}(n, As), A), \\ y &= \lfloor m/(As) \rfloor \cdot A + \mathrm{mod}(\mathrm{mod}(m, As), A), \\ c &= p \cdot s^2 + \lfloor \mathrm{mod}(m, As)/(As) \rfloor + \lfloor \mathrm{mod}(n, As)/A \rfloor, \end{aligned}$$

in which $\mathcal{AS}$ is the angular shuffling operator that rearranges the elements of an $Ah \times Aw \times s^2N$ tensor to a tensor of $AH \times AW \times N$, and $\lfloor \cdot \rfloor$ represents taking integer values. Thus, the reconstructed features maintain the same patterns as the MPI representation.

In fact, the proposed angular shuffle is a periodic version of sub-pixel convolution (Shi et al., 2016). Since we super-resolve the LF image in the macro-pixel representation, the traditional shuffle layer is no longer applicable for macro-pixel images. If using the traditional shuffle layer in our method, we should reshape the macro-pixel representation into other patterns, resulting in a more complex model. Therefore, we extend the sub-pixel convolution into the angular shuffle layer, which is specially designed for macro-pixel images.

4.3 Implementation details

In our network, convolutional layers are followed by the rectified linear unit (ReLU) activation function except for the last two layers, and we omit this function in Fig. 2 for brevity. Experimentally, we set $N = 32$ and $R = 8$. To control the model size, before the angular shuffle operation, the spatial convolution reduces the channel number by half. Our network is trained on the Y channel of LF images, so the input/output channel is 1 (i.e., $C = 1$). We use the L_1 loss to achieve better performance compared to the L_2 loss.

For the training phase, we empirically set a mini-batch size of 12 with the spatial size of 48×48 as inputs and employ the Adam optimizer with a weight decay of 10^{-4} to train our model. All weights of the layers in our network are initialized by the Xavier algorithm (Glorot and Bengio, 2010). The learning rate is initialized as 10^{-3} and reduced to one tenth of the original after 100 epochs until the validation loss converges. Our model is implemented by Pytorch on NVIDIA TITAN RTX.

5 Experiments

5.1 Datasets and settings

To validate the effectiveness of our proposed LF SR method, we conduct extensive experiments on both synthetic and real-world LF datasets. The synthetic LF images from Wanner et al. (2013) and Honauer et al. (2016) and real-world LF images from Rerabek and Ebrahimi (2016) and Kalantari et al. (2016) are collected as our dataset, which contains various LF scenes. According to the size of the dataset, we randomly select 250, 70, and 38 images as training, validation, and test datasets, respectively, without overlap. Since most existing methods choose Buddha and Mona from Wanner et al. (2013) as test samples in the HCI1 dataset, our method also divides these two images for testing on purpose. For any dataset, LF images are cropped with 7×7 angular resolution free of the border effects and then regarded as ground-truth images. Specifically, we downsample them spatially at scaling factors 2 and 4 by bicubic interpolation to acquire the LR input. The super-resolved results are evaluated by two widely used metrics, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

5.2 Ablation study

In our method, we extract both spatial and angular features from MPI representations and merge two kinds of information in integration resblocks for mutual guidance. Hence, we verify the necessities of each part by removing them separately, and the results are evaluated on the HCI2 dataset at scale 2. In our MPIN model, we deploy four integration

resblocks with 32 feature maps.

1. Spatial information

To investigate the extraction of spatial information, we remove the spatial convolution in integration resblocks and extract the angular information from the MPI representation, which is denoted as MPIN-noSpatial. For fair comparison, five integration resblocks are used to maintain the same number of parameters. Table 1 shows the results evaluated on the HCI2 dataset in terms of PSNR and SSIM. Clearly, the performance decreases severely because the intra-view spatial correlations are fully ignored.

2. Angular information

We then remove the angular convolution and the corresponding upsampler in our network to confirm their effects, and denote that model as MPIN-noAngular. Thus, the integration resblock is reduced to the simplified resblock with spatial convolution, and we employe six integration resblocks with 64 channels to keep the parameters at the same level. As illustrated in Table 1, when considering only spatial information, we observe a significant improvement compared to MPIN-noSpatial. However, in this model, it seems that we extract spatial features from SAIs and neglect inter-view angular correlations, which are widely used in the SR task.

3. Spatial-angular separation

To validate the advantage of the proposed strategy in which the spatial information and angular information are extracted in a parallel way and then merged for mutual guidance, we employ spatial-angular separable convolution (Yeung et al., 2019) in the integration resblock to acquire the spatial information and angular information alternately, referred to as MPIN-SAS in Table 1. For a fair comparison, 10 convolutional layers are deployed in one resblock with the reshaping operation. As shown in Table 1, our proposed strategy is superior to the spatial-angular separable manner.

Table 1 Investigation of different models at scale 2 (evaluated on the HCI2 dataset)

Model	PSNR (dB)	SSIM
MPIN-noSpatial	32.42	0.9420
MPIN-noAngular	35.09	0.9640
MPIN-noMG	35.27	0.9642
MPIN-SAS	35.42	0.9667
MPIN	36.04	0.9716

PSNR: peak signal-to-noise ratio; SSIM: structural similarity index

4. Mutual guidance

As spatial information and angular information are merged in the integration network for mutual guidance, 1×1 convolutional layers are eliminated in integration resblocks to demonstrate the effectiveness of the mutual guidance, which is represented as MPIN-noMG. In this way, two kinds of features interact only in the fusion stage. According to the experimental results shown in Table 1, extracting two kinds of information separately gives us a little boost. Since these features do not interact in the integration resblock, the final fusion is no longer enough to fully merge them. In our proposed MPIN, on one hand, we exploit both intra-view spatial correlations and inter-view angular correlations to make the full extraction of information. On the other hand, two kinds of features are fused progressively in integration resblocks, and thus spatial-angular correlations can be better preserved. Consequently, the angular information could benefit for the spatial SR task and the spatial information would be angular-coherent.

5. Study of N and R

We also investigate the basic network parameters, the number of integration resblocks (R) and the number of feature maps (N) in the integration resblock. As shown in Fig. 3, within a certain range, larger values of N and R would lead to better performance. This is mainly because the network becomes deeper with larger N or R. However, when the number of integration resblocks is set to 8, increasing the number of feature maps from 32 to 64, our performance is no longer improved. Similarly, when we use 32 feature maps in the integration block, increasing the number of integration resblocks from 8 to 10 does not improve the performance. This is probably because the size of the network is sufficient to fit the training dataset. In case of the limited training samples, when the network deepens to a certain extent, the model performance does not increase. Consequently, we deploy 8 integration resblocks with 32 feature maps in our proposed MPIN.

5.3 Comparisons with state-of-the-art methods

We compare the proposed MPIN with state-of-the-art methods that were primarily developed for LF SR, i.e., GB (Rossi and Frossard, 2018), RR (Farrugia et al., 2017), LFNet (Wang YL et al., 2018), LFSAS (Yeung et al., 2019), and resLF (Zhang et al.,

2019). Several methods that have been widely used for natural image SR, including bicubic interpolation (Bicubic) and EDSR (Lim B et al., 2017), are also selected for comparison; they are re-trained on the LF datasets to super-resolve each sub-aperture image separately. We calculate the quantitative index on the Y channel of the predicted image in the YCbCr space, and the Cb and Cr channels are upsampled using bicubic interpolation when generating visual results.

According to their sources, the test dataset is classified into four subsets, HCI1 from Wanner et al. (2013), HCI2 from Honauer et al. (2016), EPFL from Rerabek and Ebrahimi (2016), and StaLytro (http://lightfields.stanford.edu/LF2016.html), where there are 2, 4, 12, and 20 LF images, respectively. Table 2 demonstrates the average quantitative indexes on both synthetic and real-world datasets at magnification factors 2 and 4. Based on the results, our proposed method has significant advantages and exceeds the second-best result up to about 1.2 dB in PSNR and 0.015 in SSIM. Meanwhile, even if StaLytro contains various scenes, such as occlusions and reflective areas, the proposed method is still superior. As depicted in Table 2, due to the strong capability of a deep network, EDSR achieves a competitive performance despite the fact that it is not designed for the LF image. However, the deep learning based LF method, LFNet, performs even worse than the optimization-based method GB. The reason might be that LFNet learns the super-resolved images from horizontal and vertical stacks, resulting in insufficient integration from different directions. Even if LFSAS super-resolves the whole LF image, it alternates between two representations to extract spatial features and angular features, which means that these two kinds of features are not well integrated. Consequently, LFSAS fails in achieving considerable performance.

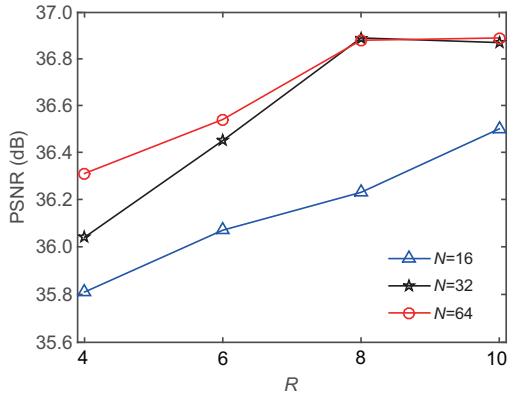


Fig. 3 Ablation investigation of the number of integration resblocks (R) and the number of feature maps (N) in the integration resblock

For visual comparisons, we display the central view of LF images reconstructed by the

Table 2 Comparison with state-of-the-art methods on the test dataset (average PSNR/SSIM for scale factors 2 and 4)

Method	Scale	PSNR (dB) / SSIM		PSNR (dB) / SSIM	
		HCI1	HCI2	EPFL	StaLytro
Bicubic	×2	37.43 / 0.9497	32.89 / 0.8903	31.53 / 0.9203	33.39 / 0.9304
RR	×2	38.13 / 0.9555	33.40 / 0.8979	32.34 / 0.9358	33.98 / 0.9456
GB	×2	39.04 / 0.9634	34.96 / 0.9278	32.98 / 0.9667	35.21 / 0.9624
LFNet	×2	38.57 / 0.9811	33.73 / 0.9544	32.74 / 0.9683	35.08 / 0.9731
LFSAS	×2	39.35 / 0.9606	34.57 / 0.9074	33.44 / 0.9376	36.87 / 0.9609
EDSR	×2	40.47 / 0.9864	35.48 / 0.9656	36.37 / 0.9762	38.81 / 0.9867
resLF	×2	40.50 / 0.9852	36.38 / 0.9764	36.04 / 0.9737	38.74 / 0.9857
MPIN (ours)	×2	**41.28 / 0.9888**	**36.89 / 0.9776**	**37.21 / 0.9779**	**39.85 / 0.9892**
Bicubic	×4	32.40 / 0.8499	28.82 / 0.7599	27.61 / 0.7973	27.99 / 0.7944
RR	×4	33.24 / 0.8702	29.40 / 0.7788	27.89 / 0.8032	28.12 / 0.8026
GB	×4	33.37 / 0.8741	29.75 / 0.7939	28.17 / 0.8186	28.69 / 0.8172
LFNet	×4	33.14 / 0.9396	29.30 / 0.8836	28.24 / 0.9017	28.75 / 0.8961
LFSAS	×4	33.87 / 0.8741	29.95 / 0.7841	28.84 / 0.8307	30.14 / 0.8508
EDSR	×4	34.55 / 0.9471	30.33 / 0.9002	29.59 / 0.9234	30.98 / 0.9316
resLF	×4	34.93 / 0.9506	30.65 / 0.9134	30.51 / 0.9224	31.09 / 0.9304
MPIN (ours)	×4	**35.56 / 0.9604**	**31.27 / 0.9208**	**30.84 / 0.9259**	**32.22 / 0.9452**

PSNR: peak signal-to-noise ratio; SSIM: structural similarity index. HCI1 and HCI2 are synthetic datasets; EPFL and StaLytro are real-world datasets. The best results are in bold

compared methods in Fig. 4 with the ground truth in the last column. Three representative images are selected at scale factor 4, which is regarded as a more difficult SR task. As the multiple views are not explored, EDSR produces ambiguous and even over-smoothed super-resolved results in Fig. 4. Although resLF is specially designed for LF images, its results are corrupted by obvious artifacts because this method uses only part of views to reconstruct the SAI. By contrast, our method could recover more texture or sharper details compared to others. Specifically, it can be observed that our method behaves well even in the occlusion area or reflective surface. By exploiting spatial-angular correlations, the proposed MPIN could reconstruct more high-frequency information during the SR process.

1. Comparison of EPIs

Since the compared methods differ in the manner of LF reconstruction, i.e., separately, orderly, or entirely, it is necessary to verify whether our method has an advantage in maintaining the inherent geometric structure in the LF image. Therefore, the super-resolved LF images are represented as EPIs horizontally and vertically, and the epipolar property is evaluated by the SSIM index. Table 3 demonstrates the results of different comparisons. Clearly, our method achieves the highest SSIM values on both synthetic and real-world datasets, which means that the proposed method can preserve more epipolar geometric structures in EPIs. To further show the advantages of our method in preserving the inherent epipolar property, we simultaneously visualize the EPI results in Fig. 4 along the blue lines in pictures. EDSR is not designed for LF SR tasks, so lines that should be slanted are vertical in the resulting images. Since SAIs are super-resolved separately in resLF, the EPIs of resLF suffer from distortion despite the high PSNR values on single images. In addition, the lines in LFSAS are consistent with the ground truth, but are relatively blurry. Compared to other methods, by taking the entire LF image into consideration, MPIN can recover more sharp lines in EPIs. Meanwhile, the spatial information is guided by the angular information via the integration resblock, which makes the reconstructions angular-coherent with the enhancement of spatial resolution.

2. Comparison of SAIs

From different angular views, we further demonstrate the reconstruction performance for each angular position. The predicted views are evaluated on the StaLytro dataset, and the PSNR values are calculated for each viewpoint and averaged on the dataset (Fig. 5). As resLF uses parts of views to process different SAIs, non-central views have relatively low quality. Nevertheless, our MPIN explores spatial-angular corrections for the entire LF image and thus achieves a better performance with a relatively balanced distribution among different views.

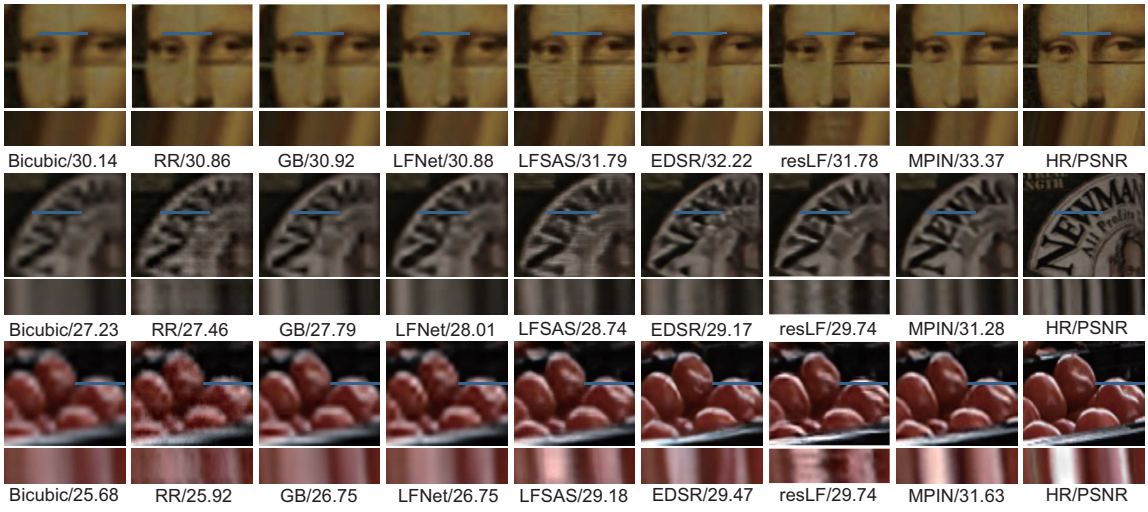


Fig. 4 Detailed ×4 super-resolution (SR) results for image Mona from HCI1, and General37 and Flowers_plants6 from StaLytro. The super-resolved central view images are visualized, and horizontal EPIs are extracted along the blue lines, where the PSNR values of central views are illustrated below. The last column shows the ground truth. References to color refer to the online version of this figure

3. Comparison of efficiency

We compare the number of parameters and running time of several learning-based methods, and Table 4 lists the results of four methods tested on the NVIDIA TITAN RTX at scale factor 2. As illustrated in Table 4, our method is slower than EDSR as all views can be processed in parallel. However, compared to resLF, because the whole LF image is processed in the macro-pixel pattern, our method is more efficient than resLF and provides higher reconstruction quality.

6 Conclusions

In this paper, we have proposed a novel integration network based on MPI representation for LF SR, called MPIN. In our method, we transformed the 4D LF data into a 2D MPI representation that couples spatial and angular information. To fully extract two kinds of information, special convolutions with subtle settings have been used to explore intra-view spatial correlations and inter-view angular correlations. To fuse them, the integration resblock has been deployed for mutual guidance, which allows our method to capture features that are more consistent with the LF image. To enlarge the MPI features, we have designed an angular shuffle layer to enhance the spatial resolution of the MPI. The qualitative and quantitative results on publicly LF datasets have demonstrated the superiority of our method over the state-of-the-art methods at different scale factors.

Table 3 Comparison of EPIs reconstructed by LF SR methods on the test dataset (average SSIM for scale factor 4)

Method	SSIM	
	Synthetic dataset	Real-world dataset
Bicubic	0.8269	0.8238
RR	0.8366	0.8249
GB	0.8485	0.8381
LFNet	0.8359	0.8363
LFSAS	0.8382	0.8524
EDSR	0.8501	0.8722
resLF	0.8694	0.8637
MPIN (ours)	**0.8756**	**0.8831**

SSIM: structural similarity index. The best results are in bold

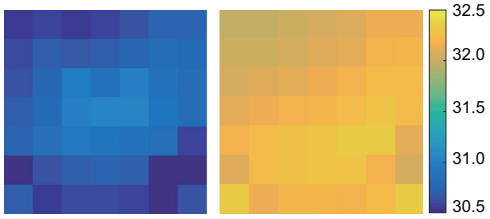


Fig. 5 Average PSNR values of SAIs reconstructed by resLF (left) and MPIN (right) at scale 4. The results are calculated on the StaLytro dataset with angular resolution of 7×7

Table 4 Comparison of the number of parameters and running time of different methods

Method	Number of parameters ($\times 10^6$)	Time (s)
LFSAS	0.81	4.25
EDSR	40.71	0.39
resLF	7.98	2.40
MPIN (ours)	1.07	1.01

Contributors

Xinya WANG and Jiayi MA designed the research. Xinya WANG and Wenjing GAO processed the data. Xinya WANG drafted the manuscript. Jiayi MA helped organize the manuscript. Jiayi MA and Junjun JIANG revised and finalized the paper.

Compliance with ethics guidelines

Xinya WANG, Jiayi MA, Wenjing GAO, and Junjun JIANG declare that they have no conflict of interest.

References

Alain M, Smolic A, 2018. Light field super-resolution via LFBM5D sparse coding. Proc 25th IEEE Int Conf on Image Processing, p.2501-2505. https://doi.org/10.1109/ICIP.2018.8451162

Bishop TE, Favaro P, 2012. The light field camera: extended depth of field, aliasing, and superresolution. *IEEE Trans Patt Anal Mach Intell*, 34(5):972-986. https://doi.org/10.1109/TPAMI.2011.168

Bishop TE, Zanetti S, Favaro P, 2009. Light field super-resolution. Proc IEEE Int Conf on Computational Photography, p.1-9. https://doi.org/10.1109/ICCPHOT.2009.5559010

Dong C, Loy CC, He KM, et al., 2014. Learning a deep convolutional network for image super-resolution. Proc 13th European Conf on Computer Vision, p.184-199. https://doi.org/10.1007/978-3-319-10593-2_13

Fan HZ, Liu D, Xiong ZW, et al., 2017. Two-stage convolutional neural network for light field super-resolution. Proc IEEE Int Conf on Image Processing, p.1167-1171. https://doi.org/10.1109/ICIP.2017.8296465

Farrugia RA, Galea C, Guillemot C, 2017. Super resolution of light field images using linear subspace projection of patch-volumes. *IEEE J Sel Top Signal Process*, 11(7):1058-1071. https://doi.org/10.1109/JSTSP.2017.2747127

Glorot X, Bengio Y, 2010. Understanding the difficulty of training deep feedforward neural networks. Proc 13th Int Conf on Artificial Intelligence and Statistics, p.249-256.

Gul MSK, Gunturk BK, 2018. Spatial and angular resolution enhancement of light fields using convolutional neural networks. *IEEE Trans Image Process*, 27(5):2146-2159. https://doi.org/10.1109/TIP.2018.2794181

Honauer K, Johannsen O, Kondermann D, et al., 2016. A dataset and evaluation methodology for depth estimation on 4D light fields. Proc 13th Asian Conf on Computer Vision, p.19-34. https://doi.org/10.1007/978-3-319-54187-7_2

Jin J, Hou JH, Chen J, et al., 2020. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2257-2266. https://doi.org/10.1109/CVPR42600.2020.00233

Kalantari NK, Wang TC, Ramamoorthi R, 2016. Learning-based view synthesis for light field cameras. *ACM Trans Graph*, 35(6):193. https://doi.org/10.1145/2980179.2980251

Levoy M, Hanrahan P, 1996. Light field rendering. Proc 23rd Annual Conf on Computer Graphics and Interactive Techniques, p.31-42. https://doi.org/10.1145/237170.237199

Li M, Diao CY, Xu DQ, et al., 2020. A non-Lambertian photometric stereo under perspective projection. *Front Inform Technol Electron Eng*, 21(8):1191-1205. https://doi.org/10.1631/FITEE.1900156

Liang CK, Ramamoorthi R, 2015. A light transport framework for lenslet light field cameras. *ACM Trans Graph*, 34(2):16. https://doi.org/10.1145/2665075

Lim B, Son S, Kim H, et al., 2017. Enhanced deep residual networks for single image super-resolution. Proc IEEE Conf on Computer Vision and Pattern Recognition Workshops, p.1132-1140. https://doi.org/10.1109/CVPRW.2017.151

Lim J, Ok H, Park B, et al., 2009. Improving the spatail resolution based on 4D light field data. Proc 16th IEEE Int Conf on Image Processing, p.1173-1176. https://doi.org/10.1109/ICIP.2009.5413719

Mitra K, Veeraraghavan A, 2012. Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition Workshops, p.22-28. https://doi.org/10.1109/CVPRW.2012.6239346

Nava FP, Luke JP, 2009. Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera. Proc 3DTV Conf: the True Vision—Capture, Transmission and Display of 3D Video, p.1-4. https://doi.org/10.1109/3DTV.2009.5069675

Ng R, Levoy M, Brédif M, et al., 2005. Light Field Photography with a Hand-Held Plenoptic Camera. Technical Report No. CTSR 2005-02, Stanford University, USA.

Peng JY, Xiong ZW, Liu D, et al., 2018. Unsupervised depth estimation from light field using a convolutional neural network. Proc Int Conf on 3D Vision, p.295-303. https://doi.org/10.1109/3DV.2018.00042

Pérez F, Pérez A, Rodríguez M, et al., 2012. Fourier slice super-resolution in plenoptic cameras. Proc IEEE Int Conf on Computational Photography, p.1-11. https://doi.org/10.1109/ICCPhot.2012.6215210

Pérez F, Pérez A, Rodríguez M, et al., 2015. Super-resolved Fourier-slice refocusing in plenoptic cameras. *J Math Imag Vis*, 52(2):200-217. https://doi.org/10.1007/s10851-014-0540-1

Rerabek M, Ebrahimi T, 2016. New light field image dataset. Proc 8th Int Conf on Quality of Multimedia Experience, p.1-2. https://doi.org/10.5281/zenodo.209499

Rossi M, Frossard P, 2017. Graph-based light field super-resolution. Proc IEEE 19th Int Workshop on Multimedia Signal Processing, p.1-6. https://doi.org/10.1109/MMSP.2017.8122224

Rossi M, Frossard P, 2018. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Trans Image Process*, 27(9):4207-4218. https://doi.org/10.1109/TIP.2018.2828983

Shi WZ, Caballero J, Huszár F, et al., 2016. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1874-1883. https://doi.org/10.1109/CVPR.2016.207

Wang TC, Zhu JY, Hiroaki E, et al., 2016. A 4D light-field dataset and CNN architectures for material recognition. Proc 14th European Conf on Computer Vision, p.121-138. https://doi.org/10.1007/978-3-319-46487-9_8

Wang YL, Hou GQ, Sun ZN, et al., 2016. A simple and robust super resolution method for light field images. Proc IEEE Int Conf on Image Processing, p.1459-1463. https://doi.org/10.1109/ICIP.2016.7532600

Wang YL, Liu F, Zhang KB, et al., 2018. LFNet: a novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Trans Image Process*, 27(9):4274-4286. https://doi.org/10.1109/TIP.2018.2834819

Wang YQ, Wang LG, Yang JG, et al., 2020. Spatial-angular interaction for light field image super-resolution. Proc 16th European Conf on Computer Vision, p.290-308. https://doi.org/10.1007/978-3-030-58592-1_18

Wanner S, Goldluecke B, 2014. Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans Patt Anal Mach Intell*, 36(3):606-619. https://doi.org/10.1109/TPAMI.2013.147

Wanner S, Meister S, Goldluecke B, 2013. Datasets and benchmarks for densely sampled 4D light fields. Proc 18th Int Workshop on Vision, Modeling, and Visualization, p.225-226. https://doi.org/10.2312/PE.VMV.VMV13.225-226

Yeung HWF, Hou JH, Chen XM, et al., 2019. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Trans Image Process*, 28(5):2319-2330. https://doi.org/10.1109/TIP.2018.2885236

Yi P, Wang ZY, Jiang K, et al., 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. Proc IEEE/CVF Int Conf on Computer Vision, p.3106-3115. https://doi.org/10.1109/ICCV.2019.00320

Yoon Y, Jeon HG, Yoo D, et al., 2015. Learning a deep convolutional network for light-field image super-resolution. Proc IEEE Int Conf on Computer Vision Workshop, p.57-65. https://doi.org/10.1109/ICCVW.2015.17

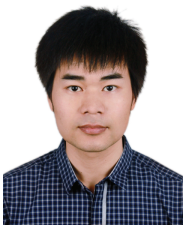
Yuan Y, Cao ZQ, Su LJ, 2018. Light-field image superresolution using a combined deep CNN based on EPI. *IEEE Signal Process Lett*, 25(9):1359-1363. https://doi.org/10.1109/LSP.2018.2856619

Yücer K, Sorkine-Hornung A, Wang O, et al., 2016. Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction. *ACM Trans Graph*, 35(3):22. https://doi.org/10.1145/2876504

Zhang S, Lin YF, Sheng H, 2019. Residual networks for light field image super-resolution. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11038-11047. https://doi.org/10.1109/CVPR.2019.01130

Zhu H, Wang Q, Yu JY, 2017. Light field imaging: models, calibrations, reconstructions, and applications. *Front Inform Technol Electron Eng*, 18(9):1236-1249. https://doi.org/10.1631/FITEE.1601727

Xinya WANG, first author of this invited paper, received her BS degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018. She is currently pursuing her PhD degree with the Electronic Information School, Wuhan University. Her research interests include neural networks, machine learning, and image processing.



Jiayi MA, corresponding author of this invited paper, received his BS degree in information and computing science and PhD degree in control science and engineering from Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a professor with the Electronic Information School, Wuhan University. He has authored or co-authored more than 200 refereed journal and conference papers. His research interests include computer vision, machine learning, and robotics. Dr. Ma has been identified in the 2020 and 2019 Highly Cited Researcher lists from Web of Science. He is an area editor of *Information Fusion*, an associate editor of *Neurocomputing*, *Sensors*, and *Entropy*, and a guest editor of *Remote Sensing*.