Neurocomputing 456 (2021) 76–87

Contents lists available at ScienceDirect

Neurocomputing

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

Learning to restore light fields under low-light imaging

Shansi Zhang *, Edmund Y. Lam

Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong

ARTICLE INFO

Article history:
Received 1 February 2021
Revised 1 April 2021
Accepted 21 May 2021
Available online 25 May 2021
Communicated by Zidong Wang

Keywords:
Light field restoration
Low-light imaging
Auxiliary feature fusion
Spatial and angular residual blocks

ABSTRACT

Light Field (LF) images have the unique advantage of recording scenes from multiple viewpoints, which provides many applications, such as refocusing and depth estimation. However, low-light conditions can severely influence these applications. In this paper, we propose a two-stage deep learning framework for the LF restoration under low-light imaging. First, there is a multi-to-one (MTO) network, which restores each view separately by utilizing multiple auxiliary views. All the views share the same feature extractor, with an efficient spatial-channel attention mechanism to extract more informative features. A channel-attention feature fusion (CAFF) module is designed to selectively fuse more useful complementary information from the auxiliary views, with a learnable global scalar to adjust the importance of the auxiliary features. Then, the outputs of the MTO network are further enhanced by an (all-to-all) ATA network, which uses spatial and angular residual blocks to process all the views synchronously for fully encoding the spatial-angular information. Extensive experiments have been conducted to demonstrate the superior performance and robustness of our method, i.e., it can restore the luminance, spatial details and angular geometries of the LF images under various light levels effectively.

1. Introduction

Different from conventional cameras, a light field (LF) camera can record both the intensities and directions of the light rays [1,2], generating 4D LF images that record the scene from multiple viewpoints. This property of LF camera enables many applications, such as post-capture refocusing [3–5], depth estimation [6–8], 3D reconstruction [9], and de-occlusion [10]. The LF images may face similar degradations with the images captured by the conventional cameras, which are either due to the physical limitations of the LF cameras or the improper light conditions. In particular, low-light environment can seriously degrade the quality of the LF images by introducing noise, color distortion and loss in detail, and therefore affect the LF applications. The commonly used solution for low-light imaging is to adopt high ISO or increase the exposure time. However, high ISO may amplify the noise level and long exposure time may cause motion blur. Our target is to develop a learning-based framework for low-light LF restoration to alleviate these problems.

As LF image can capture multiple views of the scene, exploiting its multi-view advantage is the main research direction for LF enhancement. Some methods [11–13] process each view of the LF individually by utilizing the information from some or all the views. Some methods [14–16] process all the views of the LF synchronously through one feedforward process. For the low-light image enhancement, most methods operate on the single images, mainly divided into two categories: direct image-to-image mapping approach [17–20] and illumination estimation-based approach [21–23]. The former outputs the enhanced image directly from the low-light image or raw image, and the latter contains the estimation of illumination map, which needs to be recovered to a normal light level for the image restoration. In this paper, we study on the LF restoration under low-light imaging by fully utilizing the multi-view advantage of the LF. We propose a two-stage deep learning framework, which consists of a multi-to-one (MTO) network for luminance recovery and noise suppression, and an all-to-all (ATA) network for the enhancement of spatial details and angular geometries. The main contributions of this paper are as follows:

- We synthesize raw low-light LF images with Poisson-Gaussian mixed noise at various light levels by simulating the imaging process.
- We develop a MTO network to restore each view of the LF separately by utilizing the complementary information from multiple auxiliary views. There is a shared feature extractor (FE), with an efficient spatial-channel attention mechanism for extracting more informative features for all the views. We design a

* Corresponding author at: Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong.
E-mail addresses: sszhang@eee.hku.hk (S. Zhang), elam@eee.hku.hk (E.Y. Lam).

https://doi.org/10.1016/j.neucom.2021.05.074
0925-2312/

channel-attention feature fusion (CAFF) module to selectively aggregate more useful auxiliary information, with a global scalar learned by a fully-connected (FC) layer for adjusting the importance of the auxiliary features relative to the main features.
- We develop a lightweight ATA network, which takes the outputs of the MTO network as inputs, to further enhance the spatial details and angular geometries. It adopts global residual learning and processes all the views of the LF synchronously, with spatial and angular residual blocks for fully extracting the spatial-angular information.
- Extensive experiments are conducted to demonstrate the effectiveness and robustness of our method. It can achieve promising restoration performance under various light levels.

2. Related work

2.1. LF image enhancement

LF images may suffer from similar degradations as common images, such as low resolution, low light and noise. Recently, deep-learning based methods have achieved superior performance than the traditional methods in the image enhancement tasks. Deep convolutional neural network (CNN) is also applicable to process the 4D structure of LF. Many researches focus on the LF super-resolution (SR). Some methods super-resolve one view or part of the views of the LF at each feedforward process. For example, Zhang et al. [11] proposed resLF, which adopts multiple branches to process the stacked SAIs from different angular directions for super-resolving the central view. However, this stacked approach and multi-branch architecture may be inefficient, with only partial angular information incorporated. Jin et al. [12] proposed an all-to-one network to super-resolve each view separately by utilizing all the views, and a structural consistency regularization module to preserve the parallax of the scene. However, such a network may extract much redundant information and therefore increase the computation. Moreover, Wang et al. [24] developed LFNet, which super-resolves one-row or one-column views each time by using a bi-directional recurrent CNN, with a horizontal and a vertical sub-network combined to yield the final outputs. This architecture still only utilizes limited angular information. In general, these asynchronous methods can ease the network learning and save GPU memory during training, but may not preserve the angular geometries well.

Some methods super-resolve all the views synchronously through one feedforward process. For example, Yeung et al. [14] proposed spatial-angular separable (SAS) convolution and 4D convolution for extracting both the spatial and angular features. The SAS convolution is very efficient to process LF images with good performance. Meng et al. [15] developed a high-dimensional residual network for LF reconstruction by using 4D convolutions, and they extended their work [25] by introducing generative adversarial network (GAN) to improve the performance. However, 4D convolution has a high computational cost, which may limit its applications. Wang et al. [16] developed a spatial-angular interactive network to fuse the decoupled spatial and angular features, which provides a novel approach for spatial-angular information exchange. Generally, these synchronous methods can preserve better geometric structures, but may cause learning difficulty and large GPU assumption.

LF restoration under low-light conditions can adopt the similar processing methods with LF SR. Lamba et al. [13] proposed L3Fnet, which utilizes the surrounding views to restore the central view, and integrates a global representation block to encode the LF geometry. Ge et al. [26] developed a 4D CNN architecture to recover the brightness of LF images, with a color compensation module to reduce color distortion.

2.2. Low-light image enhancement

Existing methods for low-light image enhancement with deep learning are mainly for single images. For the direct image-to-image mapping approach, some methods [17,20] operate on the raw images, i.e. Bayer pattern array, so that the networks also need to learn the demosaicing operation, which may increase the learning difficulty. Other methods take the RGB images as input. For example, Lv et al. [18] proposed MBLLEN, which adopts multiple branches to extract features at different levels, with a feature fusion module to achieve multi-branch fusion. Yang et al. [27] proposed a GAN-based method, which can learn from both the paired and unpaired data, to perform end-to-end low-light enhancement and noise reduction simultaneously.

For the illumination estimation-based methods, Wei et al. [21] proposed deep Retinex-Net, which contains a Decom-Net to decompose the input image into reflectance and illumination based on the Retinex theory, and an Enhance-Net to increase the illumination, which is then multiplied with the denoised reflectance to obtain the enhanced image. Zhang et al. [23] developed KinD, which also utilizes the Retinex theory for decomposition, and includes a reflectance restoration network to remove the degradations, and an illumination adjustment network to adjust the luminance. These decomposition-based methods introduce several sub-networks, which may increase the model complexity and training difficulty. Wang et al. [22] proposed an underexposed image enhancement method, which learns an image-to-illumination mapping. The learned illumination map is multiplied with the input image to obtain the enhanced image. However, this method is not applicable to the noisy images, as the direct multiplication with input image may amplify the noise.

3. Proposed methods

An LF image can be represented as a 4D tensor, $L \in \mathbb{R}^{U\times V\times H\times W}$, where U and V denote the angular dimensions, and H and W denote the spatial dimensions. The 4D LF can be regarded as a $U \times V$ array of sub-aperture images (SAIs), each of which has a spatial resolution of $H \times W$ and provides one view of the scene, as shown in Fig. 1(a). The SAI array can be transformed to a $UH \times VW$ macro-pixel image shown in Fig. 1(b), where the pixels with the same relative position in each SAI are put together to form a $U \times V$ macro-pixel.

As the raw LF images under low-light conditions are corrupted by severe degradations, it is a challenging task to directly restore all the views of the LF through one feedforward process. Instead, we design a two-stage framework, which consists of a MTO network and an ATA network. The former exploits the complementary information from multiple auxiliary views to restore the main view by recovering luminance and suppressing noise. The latter aims to further enhance the spatial details and geometric structures by fully encoding the spatial-angular information with synchronous approach. In what follows, we provide the details of the two networks.

3.1. Multi-to-one network

The MTO network consists of three main components: a feature extractor (FE), a channel-attention feature fusion (CAFF) module, and a decoder, as shown in Fig. 2. The main view (red-framed) is restored by utilizing the information from its 8 neighboring auxiliary views (green-framed). Let L_m denote the main view, and

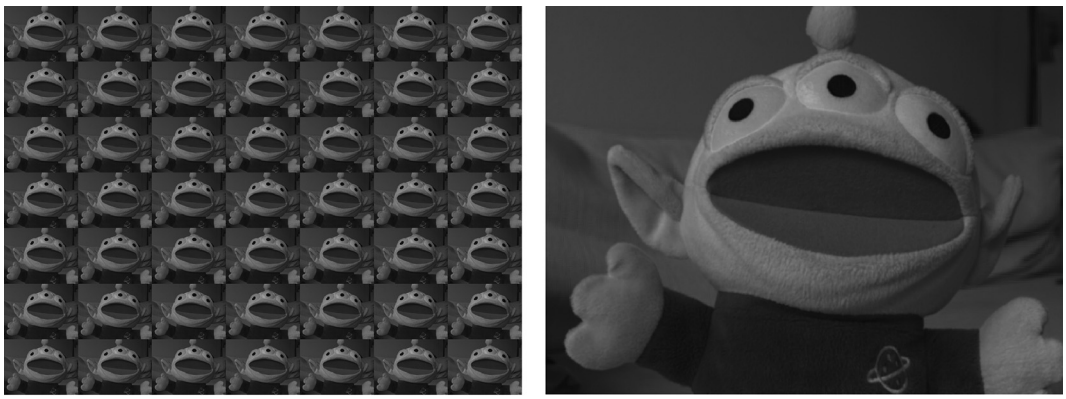

Fig. 1. SAI array pattern and Macro-pixel pattern.

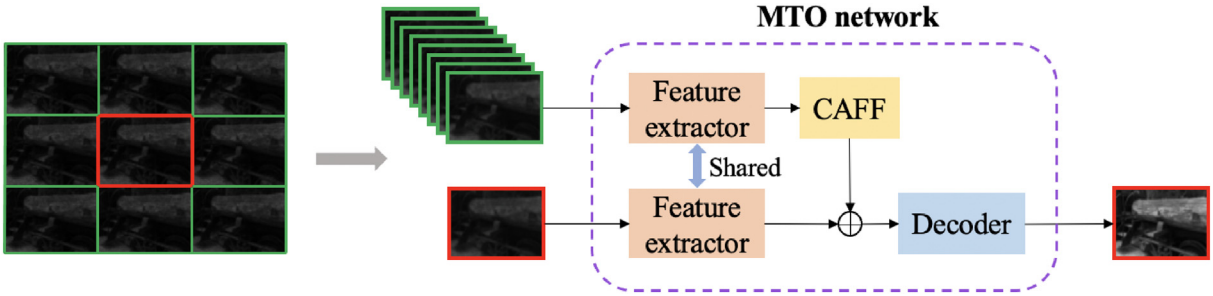

Fig. 2. The overall architecture of MTO network, which consists of a FE, a CAFF module and a decoder. It restores each view (red-framed) separately by utilizing multiple auxiliary views (green-framed).

$L_1, L_2, \cdots, L_8$ denote the auxiliary views. Thus, the restored main view L_m^{re} by the MTO network f_{MTO} is formulated as

$$L_m^{re} = f_{\text{MTO}}(L_m, L_1, L_2, \cdots, L_8). \tag{1}$$

The components of the MTO network will be introduced separately below.

3.1.1. Feature extractor

The FE, as shown in Fig. 3, is shared by all the views. Each view, including the main view and the auxiliary views, is first fed to the FE to obtain the deep features separately. Several 3×3 convolution layers with stride $= 2$ are used to gradually reduce the spatial resolution of the feature maps and increase the receptive field for capturing multi-scale information. In order to extract more informative features, we design a spatial-channel attention residual block (SCARB) following each convolution layer. Different from the serial attention in [28,29], our attention mechanism adopts parallel spatial attention and channel attention for preserving the important spatial features and channel features concurrently, and avoiding excessive suppression for some features. The detailed structure of SCARB is depicted in Fig. 3. The input features first pass through two 3×3 convolution layers to yield feature maps $M \in \mathbb{R}^{C \times H \times W}$. Given M, spatial attention is to generate a spatial attention map $m \in \mathbb{R}^{1 \times H \times W}$ by two 3×3 convolution layers and a sigmoid activation, which guarantees each attention value is between 0 and 1. The spatial attention map m is then multiplied with M to retain more important spatial features. Channel attention adopts the squeeze and excitation (SE) [30] operations to obtain the relative importance of different channels. The squeeze operation applies global average pooling (GAP) to M for encoding global information $g \in \mathbb{R}^{C \times 1 \times 1}$. The excitation operation uses two fully-connected (FC) layers followed by a sigmoid activation to derive the scaled factor $n \in \mathbb{R}^{C \times 1 \times 1}$, which is multiplied with each channel of M to preserve more informative channel features. The feature maps calibrated by the spatial attention $M \times m$ and channel attention $M \times n$ are added together to yield the enhanced features, which are added with the input features to form the residual connection.

3.1.2. Channel-attention feature fusion

Auxiliary feature fusion is important in our task, to provide effective complementary information for the restoration of the main view. The commonly used feature fusion methods include direct concatenation and summation. However, the concatenation of the features from all the auxiliary views may cause excessive GPU memory consumption, and the simple summation of the features cannot highlight more important features and suppress less useful features. Motivated by [31], we design a channel attention-based auxiliary feature fusion method, named as CAFF, to adaptively adjust the channel-wise weights for all the auxiliary features, as shown in Fig. 4. The pipeline can be formulated as

$$F_{fuse} = f_{\text{CAFF}}(f_{\text{FE}}(L_1), f_{\text{FE}}(L_2), \cdots, f_{\text{FE}}(L_8)), \tag{2}$$

where f_{CAFF} denotes the CAFF module, f_{FE} denotes the FE, and F_{fuse} is the fused auxiliary features.

All the auxiliary features $F_1, F_2, \cdots, F_8$ are first aggregated by summation: $F_s = F_1 + F_2 + \cdots + F_8$. Similar to the SE operation, GAP is applied to $F_s \in \mathbb{R}^{C \times H \times W}$ to obtain the channel-wise statistics $s \in \mathbb{R}^{C \times 1 \times 1}$, followed by a FC layer to produce a compact feature vector $s\prime \in \mathbb{R}^{C\prime \times 1 \times 1}$ $(C\prime < C)$ with fewer channels. Then, $s\prime$ is fed to different FC layers in multiple branches to generate feature vectors, each of which corresponds to one auxiliary view. Channel-wise

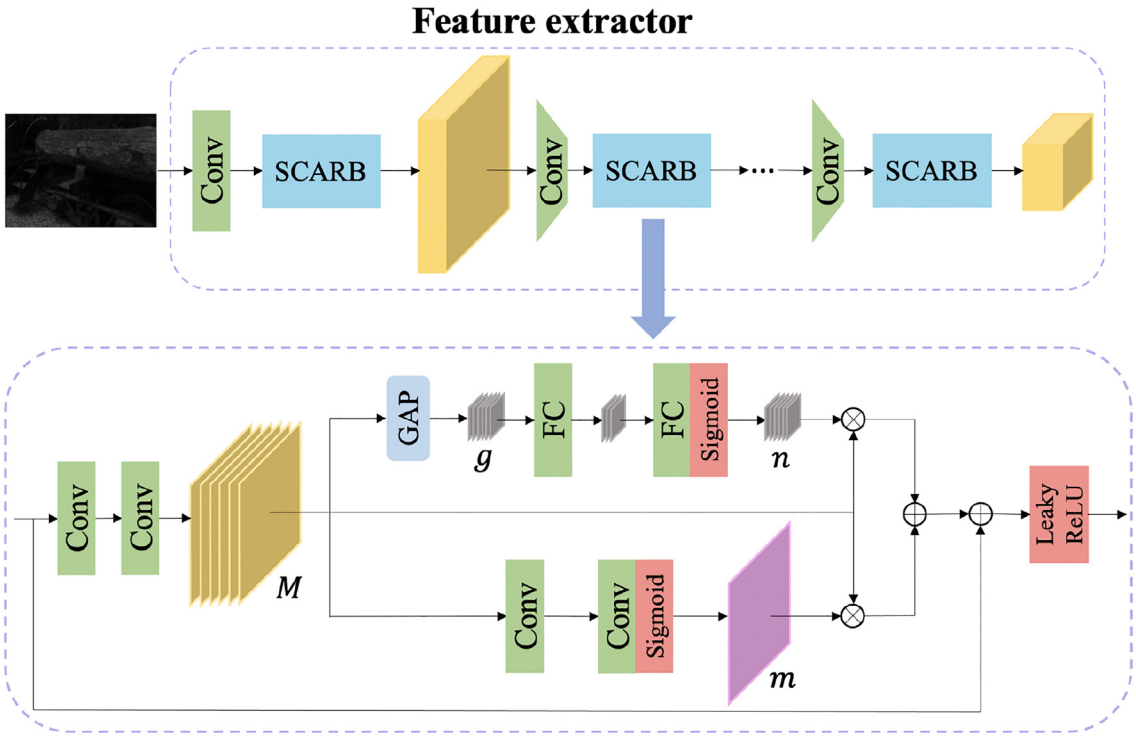

Fig. 3. The architecture of FE and SCARB. The FE consists of several convolution layers and SCARBs. The SCARB adopts parallel spatial attention and channel attention for extracting more informative features.

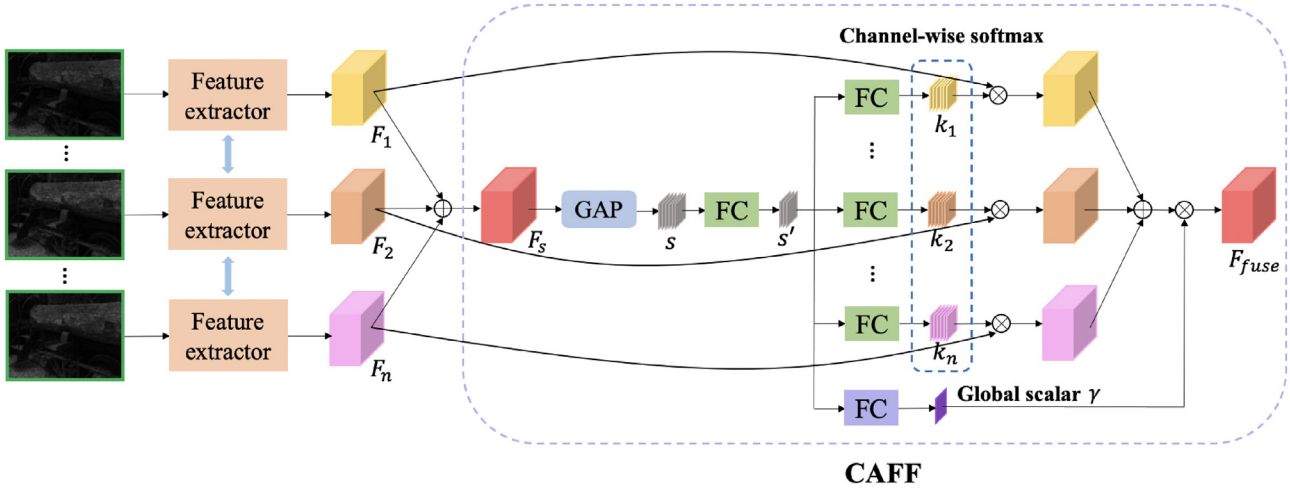

Fig. 4. The architecture of CAFF module. It learns the weights of the auxiliary features based on the channel attention, with a global scalar to adjust the importance of the auxiliary features relative to the main features.

softmax operations are then applied to these feature vectors to yield the weight vectors $k_1, k_2, \cdots, k_8$ ($k_n \in \mathbb{R}^{C\times 1\times 1}$) for each channel of the auxiliary feature maps. Another FC layer is introduced after $s\prime$ to derive a global scalar γ ($\gamma \geqslant 1$), which is multiplied with the weight vectors. Thus, the final weights for the auxiliary features become $\gamma k_1, \gamma k_2, \cdots, \gamma k_8$. This global scalar can adaptively adjust the importance of the auxiliary features relative to the main features while maintaining the relative importance among the auxiliary features. γk_n is clipped to $[0, 1]$ to guarantee the weights of the auxiliary features does not exceed the weights of the main features. Hence, the fused auxiliary features are obtained by

$$F_{fuse} = F_1 \times \gamma k_1 + F_2 \times \gamma k_2 + \cdots + F_8 \times \gamma k_8 \tag{3}$$

3.1.3. Encoder-decoder architecture

After introducing the auxiliary feature fusion, we can build the encoder-decoder architecture for restoring the main view, as shown in Fig. 5. The FE can be treated as the encoder to extract multi-scale contextual information. The extracted features of the main view are then added with the fused auxiliary features by the CAFF module, followed by a convolution layer for further feature aggregation. The decoder is to gradually recover the feature maps to the input resolution through several bilinear upsampling operations. Skip connections are used to concatenate the upsampled features in the decoder and the corresponding features in the encoder, followed by the residual blocks, to fuse the features with high-level semantic information and high resolution for more

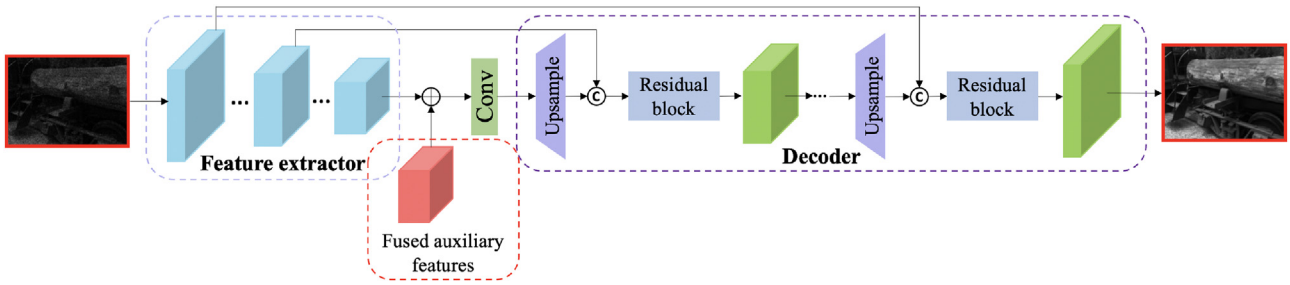

Fig. 5. The encoder-decoder architecture for restoring the main view. The fused auxiliary features are added to the main features extracted by the FE. The decoder gradually recovers the spatial resolution of the feature maps through bilinear upsampling, with skip connections to fuse the features with high-level semantic information and high resolution.

precise restoration. The restored main view by this encoder-decoder architecture can be formulated as

$$L_m^{re} = f_{\mathrm{DE}}\left(f_{\mathrm{c}}\left(f_{\mathrm{FE}}(L_m) + F_{fuse}\right)\right), \tag{4}$$

where f_{c} denotes the convolution layer after the summation of the main features and the fused auxiliary features, and f_{DE} denotes the decoder.

3.2. All-to-all network

In order to further improve the restoration performance, we design a lightweight ATA network shown in Fig. 6, which takes the outputs of the MTO network as inputs, and enhances all the views synchronously through one feedforward process.

The information flow of the ATA network is described as follows. The LF image output from the MTO network with SAI array pattern $L^{re} \in \mathbb{R}^{UV\times C\times H\times W}$ first pass through a 2D spatial convolution layer to extract spatial features for each SAI. Then, several spatial-angular modules, each of which consists of a spatial residual block and an angular residual block, are used to fully extract the spatial and angular features. Each spatial residual block contains two 3×3 spatial convolution layers and each angular residual block contains two 3×3 angular convolution layers. Besides the convolution operation, the first convolution layer includes group normalization [32] and leaky ReLU activation, and the second convolution layer only includes group normalization, the output of which is added with the input followed by a leaky ReLU activation. When extracting the angular features, the feature maps with SAI array pattern need to be converted to the macro-pixel pattern $F_{mp} \in \mathbb{R}^{HW\times C\times U\times V}$, and then 2D angular convolutions are applied to operate on every macro-pixel. After the angular residual block, the macro-pixel pattern is converted to the SAI array pattern again, followed by the next spatial-angular module. The output of the last spatial-angular module is fed to the final spatial convolution layer to produce the residual output, which is added to the input SAI array to yield the final restoration result. This global residual learning can ease the network learning by concentrating on the high-frequency details. The final enhanced LF can be formulated as

$$L^{en} = L^{re} + f_{\mathrm{ATA}}(L^{re}), \tag{5}$$

where L^{re} is the restored LF from the MTO network and f_{ATA} denotes the ATA network.

3.3. Loss function

The loss function used for training the MTO and ATA networks is a mixture of MSE loss, structural similarity (SSIM) [33] loss and perceptual loss [34] between the restored results and ground truths. The loss of the ATA network is calculated by averaging all the views.

The MSE loss is defined as

$$\ell_{MSE} = \left\| L^{en} - \widehat{L} \right\|_2^2, \tag{6}$$

where $\widehat{L}$ is the ground truth LF image.

The SSIM loss between any two images X and Y is defined as

$$\ell_{SSIM}(X,Y) = 1 - \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)}, \tag{7}$$

where μ_X and μ_Y are the averages of pixel values, σ_X^2 and σ_Y^2 are the variances, σ_{XY} is the covariance, and c_1 and c_2 are two constants. Thus, the SSIM loss between the restored LF and ground truth is expressed as $\ell_{SSIM}\left(L^{en}, \widehat{L}\right)$.

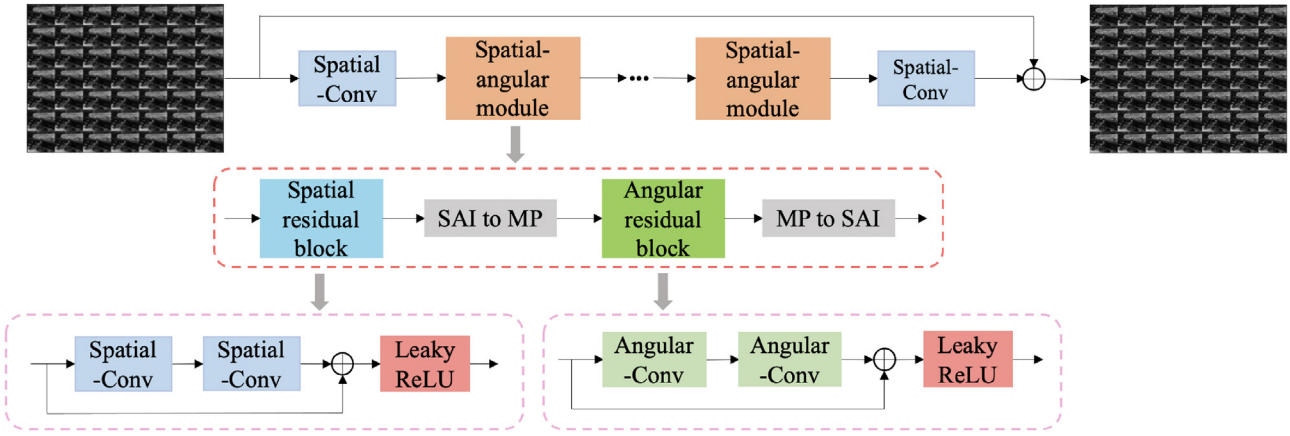

Fig. 6. The architecture of ATA Network. It mainly consists of spatial residual blocks and angular residual blocks to extract spatial and angular features alternately, and adopts global residual learning to focus on the high-frequency details.

The perceptual loss is defined as the difference of the features extracted by the pretrained VGG-19 [35] network

$$\ell_{percept} = \left\| \phi_{ij}(L^{en}) - \phi_{ij}\left(\widehat{L}\right) \right\|_2^2, \tag{8}$$

where ϕ_{ij} means the features after the jth convolution and before the ith maxpooling layer.

The overall loss function is the weighted sum of the three loss terms

$$\ell_{total} = w_1 \ell_{MSE} + w_2 \ell_{SSIM} + w_3 \ell_{percept}, \tag{9}$$

where w_1, w_2 and w_3 are their corresponding weights.

4. Experiments

In this section, we first introduce the synthetic method of low-light LF images and then the implementation details. After that, we perform ablation study and comparison experiments to verify the effectiveness of our method.

4.1. Synthesis of raw low-light LF images

Image formation is based on the Poisson process that describes the arrival of photons from scene to sensor [36]. By considering various distortions, the observed image can be represented as:

$$x_{\text{observe}} = \text{ADC}[\alpha \cdot \text{Poisson}(x_{\text{scene}} + n_d) + n_r]. \tag{10}$$

where n_d is the dark current due to the random generation of electrons and holes within the depletion region, n_r is the readout noise caused by the electron readout and voltage amplification, α is the sensor gain to obtain voltage at each pixel, and ADC is to convert the voltages to the pixel values. The photon arrival follows a Poisson distribution, which results in the shot noise. The dark current follows a Poisson distribution and the readout noise follows a Gaussian distribution.

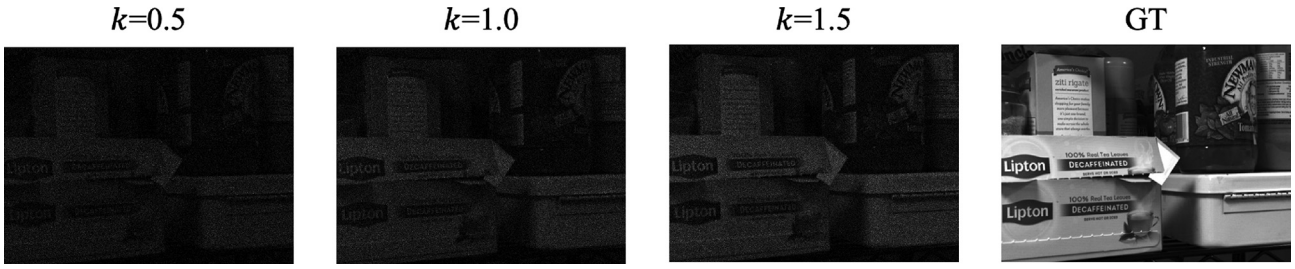

Fig. 7. Central views of the synthetic low-light LFs at $k = 0.5, 1.0, 1.5$.

Table 1
Quantitative results of different configurations for the MTO network on the test sets at $k = 0.5, 1.0, 1.5$.

Model	Light level	PSNR↑	SSIM↑	LPIPS↓
w/o CAFF	0.5	22.68	0.6475	0.2236
Direct summation	0.5	23.90	0.6807	0.1981
w/o global scalar	0.5	23.64	0.6799	0.2103
w/o SCA	0.5	24.08	0.6861	0.1975
Serial SCA	0.5	23.91	0.6865	0.2014
Our MTO	0.5	**24.11**	**0.6947**	**0.1912**
w/o CAFF	1.0	24.10	0.7085	0.1883
Direct summation	1.0	25.28	0.7390	0.1607
w/o global scalar	1.0	25.16	0.7382	0.1695
w/o SCA	1.0	25.41	0.7428	0.1591
Serial SCA	1.0	25.27	0.7423	0.1618
Our MTO	1.0	**25.56**	**0.7508**	**0.1581**
w/o CAFF	1.5	24.62	0.7368	0.1644
Direct summation	1.5	25.90	0.7664	0.1415
w/o global scalar	1.5	26.01	0.7681	0.1471
w/o SCA	1.5	26.06	0.7701	0.1398
Serial SCA	1.5	25.99	0.7686	0.1406
Our MTO	1.5	**26.10**	**0.7752**	**0.1383**

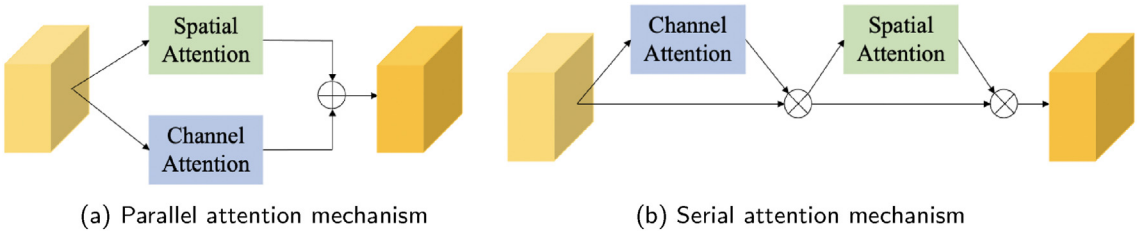

Fig. 8. Parallel and serial attention mechanisms.

In our simulation, each ground truth LF image is first divided by its mean pixel value and multiplied by a factor k. Then, Poisson process is applied to obtain the photon count at each pixel, and a readout noise with standard deviation of $0.25e^{-}$ is added. After that, the pixel values are scaled to $0 \sim 255$ to yield the raw low-light LF images, which suffer from the Poisson-Gaussian mixed noise. We ignore the dark current as it is usually very small compared with other noises. Here, the factor k is used to control the light levels, and can also be interpreted as the photons per pixel (ppp) [37], which is the average number of photons each pixel detects during exposure. We synthesize the low-light LF images by using $k = 0.5, 1.0, 1.5$ based on the LF dataset from Kalantari [38] and Stanford Lytro LF Archive [39].

We select 100 images for training and 30 images for testing from the Kalantari dataset, and 200 images for training and 50 images for testing from the Stanford Lytro dataset. Each LF image can synthesize multiple low-light LF images with different k values. We only retain the central 7×7 SAIs for all the LF images. Fig. 7 shows the central views of the synthetic low-light LFs at $k = 0.5, 1.0, 1.5$. As can be seen, the synthetic low-light images are corrupted by noise, and the degradations become more severe with the decrease of k.

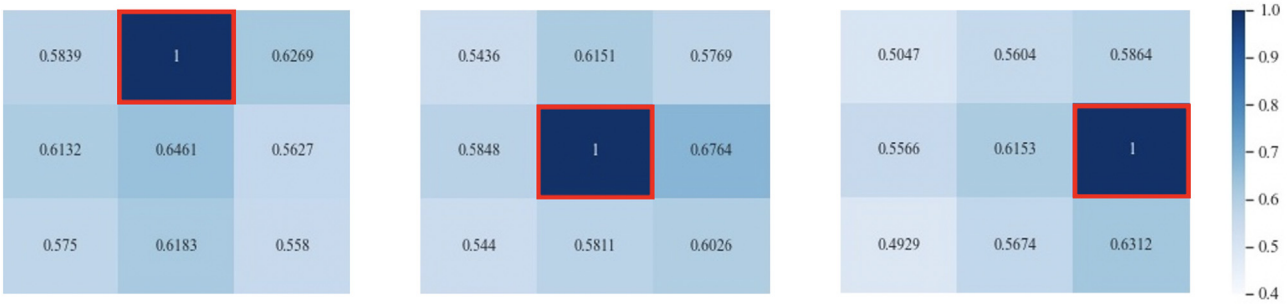

Fig. 9. The average weights of auxiliary views. The main views are red-framed and has weight of 1.

Table 2
Quantitative results of our MTO and ATA network on the test sets at $k = 0.5, 1.0, 1.5$.

Model	Light level	PSNR↑	SSIM↑	LPIPS↓
Our MTO + ATA	0.5	26.78	0.8126	0.1223
	1.0	27.81	0.8519	0.1010
	1.5	28.13	0.8682	0.0909

Table 3
Params. and FLOPs of different configurations.

Model	Params.	FLOPs
w/o CAFF	2.79 M	4.15G
Direct summation	3.94 M	19.83G
w/o global scalar	4.01 M	19.83G
w/o SCA	3.90 M	18.94G
Serial SCA	4.01 M	19.83G
Our MTO	4.01 M	19.83G
Our MTO + ATA	4.60 M	138.83G

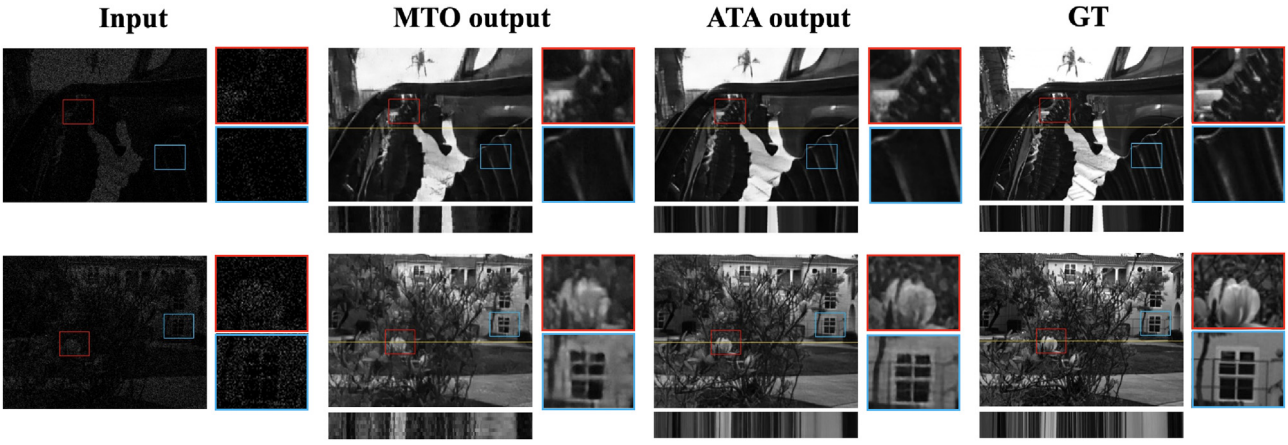

Fig. 10. Visual results of our MTO and ATA networks at $k = 0.5$.

4.2. Implementation details

In the MTO network, one 3×3 convolution layer with *stride* $= 1$ is first used for initial feature extraction, and four 3×3 convolution layers with *stride* $= 2$ are used for down-sampling. Each convolution layer is followed by a SCARB, so there are five SCARBs in the FE, where the resolution of the feature maps is gradually reduced by 16 times and the number of channels is increased from 16 to 256. Each convolution layer is the combination of convolution, group normalization with group number of 8 and leaky ReLU activation. Considering the GPU memory capacity while avoiding extracting too many redundant features, we only choose the nearest 3×3 SAIs around the main view as the auxiliary views. The MTO network, including the FE, CAFF and decoder, can be trained end-to-end. For the ATA network, four spatial-angular modules with 64 channels are used to build a lightweight architecture.

During the training of MTO network, each input low-light LF was synthesized from the ground truth by a random k value, and the SAIs of LFs were randomly cropped to 256×256. The main view was selected from each LF randomly. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was initially set to 1×10^{-3} and decreased by a rate of 0.8 every 50 epochs. The

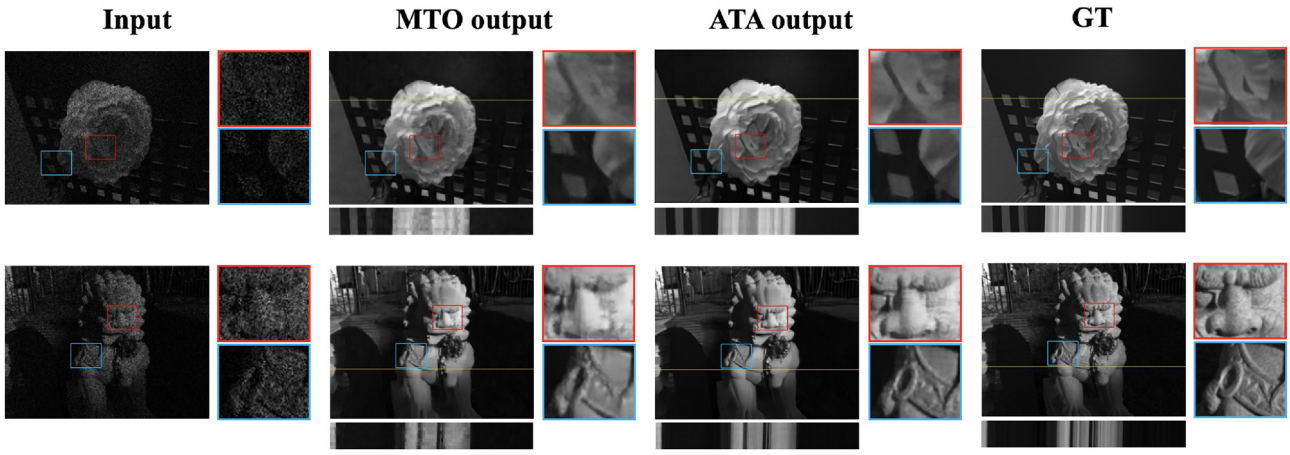

Fig. 11. Visual results of our MTO and ATA networks at $k = 1.0$.

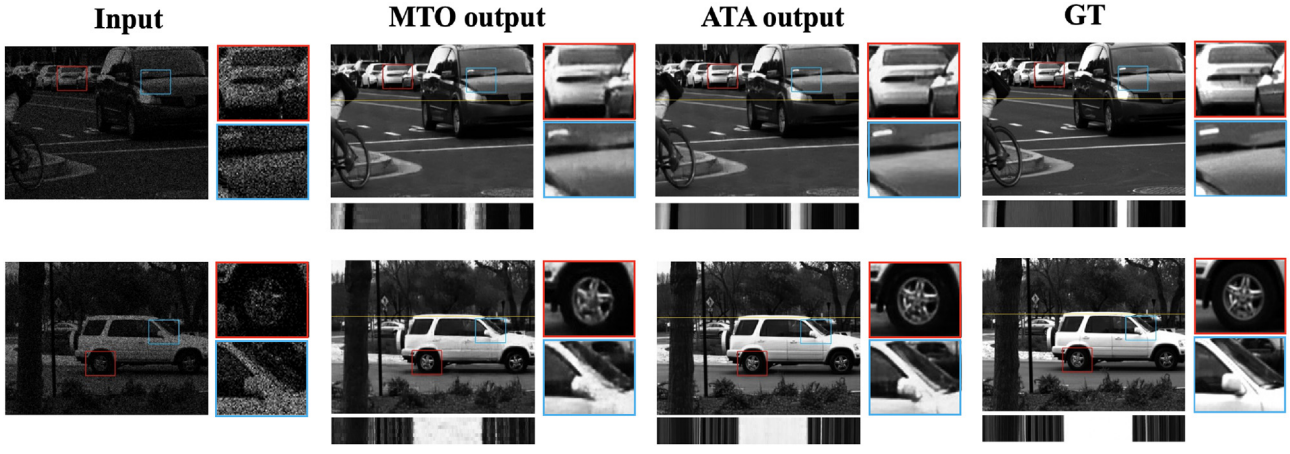

Fig. 12. Visual results of our MTO and ATA networks at $k = 1.5$.

Table 4
Quantitative results of different methods on the test sets at various light levels.

Model	Light level	PSNR↑	SSIM↑	LPIPS↓	NIQE↓
MBLLEN [18]	0.5	23.01	0.6559	0.2112	7.080
Improved LFSAS [14]	0.5	21.20	0.7373	0.1767	6.327
Ours	0.5	**26.78**	**0.8126**	**0.1223**	**5.298**
MBLLEN [18]	1.0	24.27	0.7155	0.1745	6.904
Improved LFSAS [14]	1.0	22.31	0.7709	0.1438	6.172
Ours	1.0	**27.81**	**0.8519**	**0.1010**	**4.911**
MBLLEN [18]	1.5	24.75	0.7437	0.1514	6.762
Improved LFSAS [14]	1.5	22.93	0.7835	0.1306	6.158
Ours	1.5	**28.13**	**0.8682**	**0.0909**	**4.766**

MTO network was trained for 300 epoch with batch size set to 6. Then we used the trained MTO network to generate the preliminary restoration results for all the low-light LFs at $k = 0.5, 1.0, 1.5$, so each ground truth corresponded to three intermediate results. During the training of ATA network, one of the three intermediate results was chosen as the input and all the SAIs were randomly cropped to 128×128. We used Adam optimizer with a initial learning rate of 1×10^{-4} and a decay rate of 0.8 every 40 epochs. The ATA network was trained for 100 epochs with batch size set to 1. Moreover, we used data augmentation of random flipping during the training of the two networks. The weights of MSE loss, SSIM loss and perceptual loss, which is defined by ϕ_{54}, were set to $10, 1$ and 5, respectively. Our method was implemented by PyTorch on Nvidia Tesla V100-SXM2 16 GB GPU. It took about three days for training the two networks.

4.3. Ablation study

In this subsection, we investigate the effectiveness of our proposed modules by comparing the performance of different configurations on the test sets at $k = 0.5, 1.0, 1.5$. We use PSNR and SSIM, along with learned perceptual image patch similarity (LPIPS) [40] as the quantitative metrics to evaluate the performance, which are calculated by averaging all the LF views over the whole test sets. For the PSNR and SSIM, higher value is better, while for the LPIPS, lower value is better.

Firstly, we study the benefit of our CAFF module by removing it from the MTO network. Thus, the network does not utilize any auxiliary view to incorporate the complementary information, which is equivalent to the single image restoration. As shown in Table 1, the configuration 'w/o CAFF' obtains much lower PSNR and SSIM, and higher LPIPS than 'Our MTO' at various light levels, which indicates that auxiliary feature fusion can significantly boost the restoration performance. Moreover, we replace the CAFF module with direct summation of auxiliary features ('Direct summation'), which means all the auxiliary features have the same importance. As can be seen, it has worse quantitative results than 'Our MTO', which may be because the direct summation can not highlight more informative auxiliary features and even weaken the effect of main features. Furthermore, we remove the global scalar from the CAFF module ('w/o global scalar'), and the resulting worse quantitative results indicates that the reduced weights of auxiliary features may also degrade the performance.

Next, we investigate the advantage of our parallel spatial-channel attention mechanism. We first replace SCARBs with simple residual blocks ('w/o SCA'), which obtains worse quantitative results than 'Our MTO'. Then we compare the performance of parallel and serial attention mechanisms, the structures of which are depicted in Fig. 8. In the serial attention mechanism ('Serial SCA'), the feature maps are calibrated by the channel attention and spatial attention sequentially. It can be seen that the quantitative results of 'Serial SCA' are also worse than 'Our MTO', which may be because the serial attention weakens some features exces-

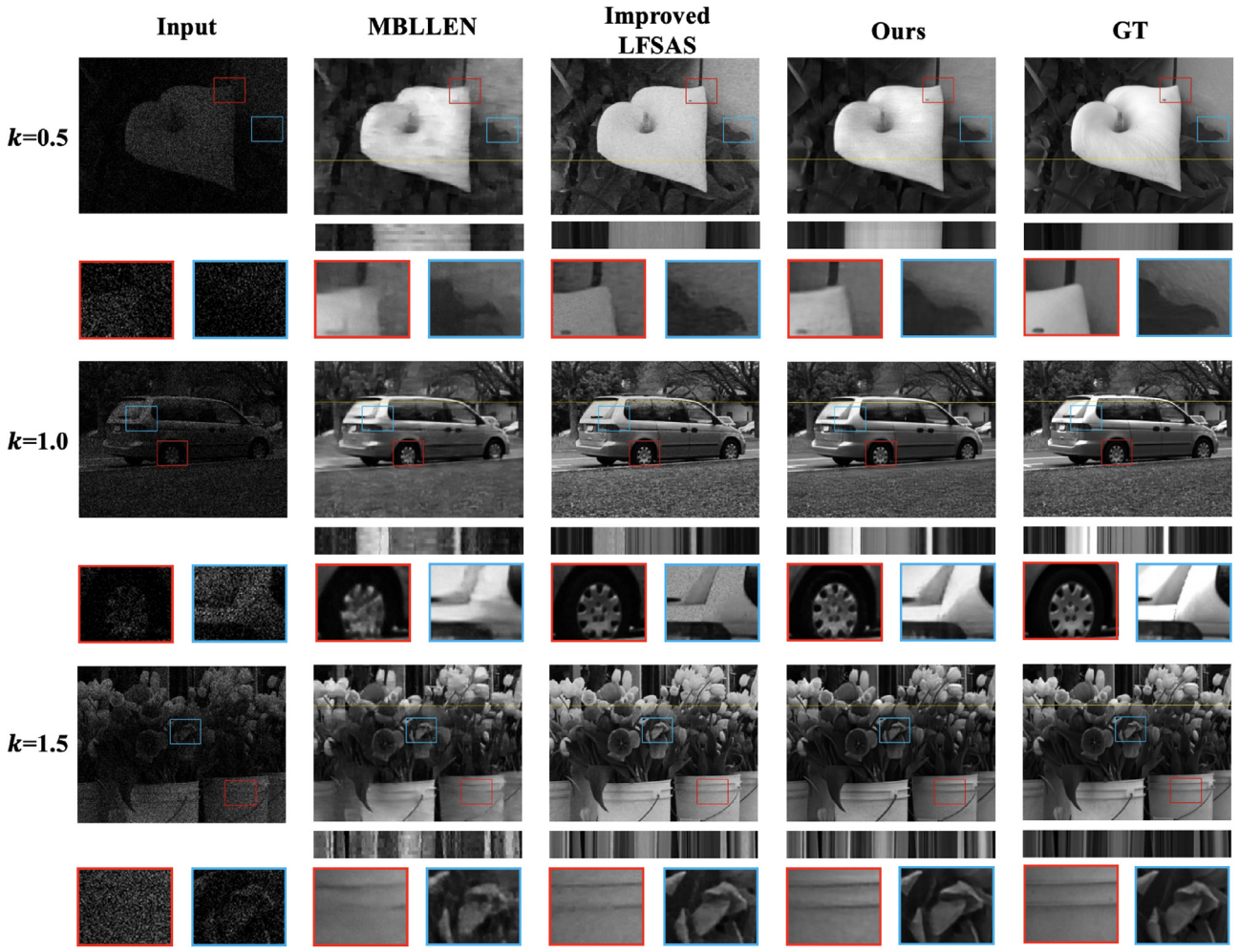

Fig. 13. Visual results of different methods at $k = 0.5, 1.0, 1.5$.

sively by successive suppression. Therefore, our parallel spatial-channel attention mechanism is more conducive to the important feature extraction.

To check the learning results of CAFF, we visualize the weights of auxiliary views corresponding to three main views, as shown in Fig. 9. The main views are red-framed with weight of 1. The first main view is in the first row of the SAI array and the third main view is in the last column of the SAI array, so they are not in the center of the 3×3 SAIs. The weights of auxiliary views are calculated by averaging along all the 256 channels, which therefore results in the approximate values. It still can be seen that the auxiliary views close to the main view tend to have larger weights than the others, demonstrating the effect of adaptive learning for the auxiliary weights.

Finally, we verify the effectiveness of our ATA network by using the complete two-stage pipeline. As shown in Table 2, by introducing ATA network, PSNR is further improved, SSIM has significant increase, and LPIPS has obvious decrease. Taking the MTO outputs as inputs, the ATA network can learn more easily by focusing on the high-frequency details and angular geometries. Therefore, our two-stage learning framework is very effective for better restoration performance.

The number of parameters (Params.) and FLOPs of different configurations are recorded in Table 3. The FLOPs are calculated according to the input size during training. Our MTO and ATA networks have 4.60M parameters in total. Specifically, the ATA network is very lightweight with only 0.59M parameters. However, it results in a significant increase of FLOPs, as it processes all the views of LF synchronously with full resolution throughout the feedforward process. Params. and FLOPs can be largely reduced by excluding the auxiliary features at the expense of obvious performance degradation. Moreover, our spatial-channel attention mechanism only introduces additional 0.11M parameters, and the other configurations for the MTO network have approximate Params. and FLOPs with 'Our MTO' but with worse performance.

Some visual results of our MTO and ATA networks on the test data at different light levels are shown in Fig. 10, 11 ∼ Fig. 12. Only the central views of LFs are presented, and two color-framed patches are zoomed in for better visualization. The yellow lines are the positions where epipolar-plane images (EPIs) are located. Each EPI is obtained by fixing one spatial and one angular coordinate of a LF image, which can indicate the LF parallax. It can be seen that the MTO network can effectively restore the illumination and suppress the noise, but some objects are still a little blurry and the angular geometries are not preserved well. After refinement by the ATA network, more spatial details and geometric structures are recovered, which are attributed to the residual learning and spatial-angular feature extraction. Therefore, our two-stage learning framework is capable of restoring the LFs under different light levels, even the extreme low-light conditions, which verifies its effectiveness and robustness.

4.4. Comparison with other methods

In this section, we make comparison with other methods related to the low-light and LF enhancement. The networks of other methods are re-trained by our dataset. As we use the direct image-to-image mapping approach for low-light enhancement, we choose MBLLEN [18], which adopts the same manner with us, as the comparative object. To further verify the effectiveness of our two-stage method, we compare with an improved version of LFSAS [14], which restores all the views of LF synchronously by one stage. As LFSAS is designed for super-resolution, we remove the transposed convolution for adapting to our task. We also replace its original spatial and angular convolution layers with our spatial and angular residual blocks for better performance.

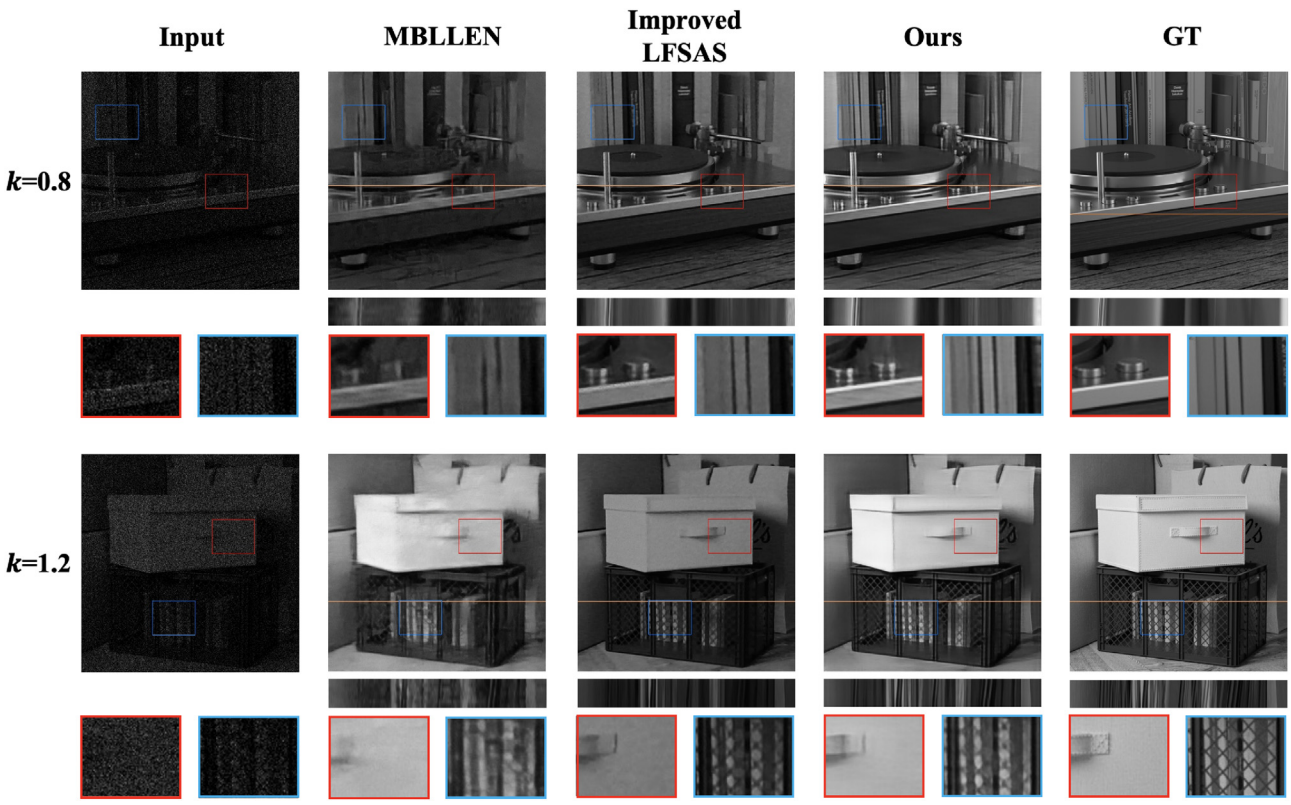

Fig. 14. Visual results of different methods on the other dataset at $k = 0.8$ and $k = 1.2$.

The quantitative results of different methods on the test sets are recorded in Table 4. In addition to the PSNR, SSIM and LPIPS, we also adopt Natural Image Quality Evaluator (NIQE) [41], a no-reference image quality assessment method, to evaluate the restoration results, with lower value meaning better. It can be seen that our method achieves much higher PSNR and SSIM, and lower LPIPS and NIQE than other methods. Some comparative visual results at various light levels are shown in Fig. 13. As can be seen, the outputs of MBLLEN can not restore some object details and the EPIs have confused structures. As MBLLEN is based on one-to-one restoration without utilizing any angular information, it can not recover the geometric structures effectively. The outputs of the improved LFSAS can preserve relatively better spatial details and geometric structures. However, the restored images lack smoothness and still contain noise, which results in low PSNR. In comparison, our method can obtain more natural LF images with fine spatial details and angular geometries.

We also test these methods on the HCI dataset [42], which consists of synthetic LF images, different from the real scenes in the Kalantari and Stanford datasets that we used for training. Fig. 14 presents two visual results of different methods at $k = 0.8$ and $k = 1.2$, which are different from the light levels used during training. Again, it can be seen that our method can restore the luminance, spatial details and angular geometries better than the other methods, further demonstrating its good performance and robustness.

5. Conclusion

In this paper, we propose a two-stage learning framework for low-light LF restoration. The first stage is performed by the MTO network which aims to recover luminance and suppress noise by utilizing multiple auxiliary views. An efficient spatial-channel attention mechanism is designed for extracting more informative features. The auxiliary features are selectively fused by the CAFF module, with a learned global scalar to adjust the importance of auxiliary features with respect to the main features. In the second stage, the outputs of the MTO network are input to the ATA network, which aims to further enhance the spatial details and angular geometries by fully extracting the spatial-angular information. The ATA network is very lightweight, but can improve the restoration results significantly. Extensive experiments have been conducted to demonstrate the superior performance and robustness of our method.

CRediT authorship contribution statement

Shansi Zhang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft. **Edmund Y. Lam:** Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgments

The work is supported in part by the Research Grants Council of Hong Kong (GRF 17201818, 17200019, 17201620) and the University of Hong Kong (104005864).

References

[1] N. Chen, Z. Ren, D. Li, E.Y. Lam, Analysis of the noise in backprojection light field acquisition and its optimization, Appl. Opt. 56 (2017) F20–F26.
[2] H. Duan, L. Mei, J. Wang, L. Song, N. Liu, A new imaging model of Lytro light field camera and its calibration, Neurocomputing 328 (2019) 189–194.
[3] J. Fiss, B. Curless, R. Szeliski, Refocusing plenoptic images using depth-adaptive splatting, in: IEEE International Conference on Computational Photography, pp. 1–9..
[4] X. Zhang, Y. Wang, J. Zhang, L. Hu, M. Wang, Light field saliency vs. 2D saliency: A comparative study, Neurocomputing 166 (2015) 389–396.
[5] Y. Wang, J. Yang, Y. Guo, C. Xiao, W. An, Selective Light Field Refocusing for Camera Arrays Using Bokeh Rendering and Superresolution, IEEE Signal Process. Lett. 26 (2019) 204–208.
[6] T. Wang, A.A. Efros, R. Ramamoorthi, Depth Estimation with Occlusion Modeling Using Light-Field Cameras, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 2170–2181.
[7] C. Shin, H.G. Jeon, Y. Yoon, I.S. Kweon, S.J. Kim, EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4748–4757..
[8] F. Liu, G. Hou, Z. Sun, T. Tan, High quality depth map estimation of object surface from light-field images, Neurocomputing 252 (2019) 3–16.
[9] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, M.H. Gross, Scene reconstruction from high spatio-angular resolution light field, ACM Trans. Graphics 32 (2013).
[10] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, Y. Guo, DeOccNet: Learning to See Through Foreground Occlusions in Light Fields, in: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 118–127..
[11] S. Zhang, Y. Lin, H. Sheng, Residual networks for light field image super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11046–11055..
[12] J. Jin, J. Hou, J. Chen, S. Kwong, Light Field Spatial Super-Resolution via Deep Combinatorial Geometry Embedding and Structural Consistency Regularization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2260–2269..
[13] M. Lamba, K. Kumar, K. Mitra, Harnessing multi-view perspective of light fields for low-light imaging, IEEE Trans. Image Process. 30 (2021) 1501–1513.
[14] H.W.F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, Y.Y. Chung, Light Field Spatial Super-Resolution Using Deep Efficient Spatial-Angular Separable Convolution, IEEE Trans. Image Process. 28 (2019) 2319–2330.
[15] N. Meng, H.K.-H. So, X. Sun, E.Y. Lam, High-dimensional dense residual convolutional neural network for light field reconstruction, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 873–886.
[16] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, Y. Guo, Spatial-Angular Interaction for Light Field Image Super-Resolution, in: European Conference on Computer Vision (ECCV), pp. 290–308..
[17] C. Chen, Q. Chen, J. Xu, V. Koltun, Learning to see in the dark, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3291–3300..
[18] F. Lv, F. Lu, J. Wu, C. Lim, MBLLEn: Low-light image/video enhancement using CNNs, in: British Machine Vision Conference (BMVC)..
[19] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, Z. Wang, EnlightenGAN: Deep light enhancement without paired supervision, IEEE Trans. Image Process. 30 (2021) 2340–2349.
[20] E. Schwartz, R. Giryes, A.M. Bronstein, DeepISP: Toward Learning an End-to-End Image Processing Pipeline, IEEE Trans. Image Process. 28 (2019) 2170–2181.
[21] C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement, in: British Machine Vision Conference (BMVC)..
[22] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, J. Jia, Underexposed photo enhancement using deep illumination estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6849–6857..
[23] Y. Zhang, J. Zhang, X. Guo, Kindling the darkness: A practical low-light image enhancer, in: ACM International Conference on Multimedia, pp. 1632–1640..
[24] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, T. Tan, LFNet: A Novel Bidirectional Recurrent Convolutional Neural Network for Light-Field Image Super-Resolution, IEEE Trans. Image Process. 27 (2018) 4274–4286.
[25] N. Meng, Z. Ge, T. Zeng, E.Y. Lam, LightGAN: A Deep Generative Model for Light Field Reconstruction, IEEE Access 8 (2020) 116052–116063.
[26] Z. Ge, L. Song, E.Y. Lam, Light field image restoration in low-light environment, in: Future Sensing Technologies, volume 11525 of Proceedings of the SPIE, p. 115251H..
[27] Q. Yang, Y. Wu, D. Cao, M. Luo, T. Wei, A lowlight image enhancement method learning from both paired and unpaired data by adversarial training, Neurocomputing 433 (2021) 83–95.
[28] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: European Conference on Computer Vision (ECCV), pp. 3–19..
[29] L. He, X. Gong, S. Zhang, L. Wang, F. Li, Efficient attention based deep fusion CNN for smoke detection in fog environment, Neurocomputing 434 (2021) 224–238.
[30] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141..
[31] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 510–519..

[32] Y. Wu, K. He, Group Normalization, in: European Conference on Computer Vision (ECCV), pp. 3–19..
[33] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.
[34] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690..
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint (2014)..
[36] W. Wang, X. Chen, C. Yang, X. Li, X. Hu, T. Yue, Enhancing Low Light Videos by Exploring High Sensitivity Camera Noise, in: IEEE International Conference on Computer Vision (ICCV), pp. 4111–4119..
[37] Y. Chi, A. Gnanasambandam, V. Koltun, S.H. Chan, Dynamic Low-light Imaging with Quanta Image Sensors, in: European Conference on Computer Vision (ECCV), pp. 122–138..
[38] N.K. Kalantari, T.C. Wang, R. Ramamoorthi, Learning-based view synthesis for light field cameras, ACM Trans. Graphics 35 (2016).
[39] R. Shah, G. Wetzstein, A.S. Raj, M. Lowney, Stanford lytro light field archive (2016)..
[40] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595..
[41] A. Mittal, R. Soundararajan, A.C. Bovik, Making a Completely Blind Image Quality Analyzer, IEEE Signal Process. Lett. 20 (2013) 209–212.
[42] K. Honauer, O. Johannsen, D. Kondermann, B. Goldluecke, A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields, in: Asian Conference on Computer Vision (ACCV), pp. 19–34..

Shansi Zhang received the B.S. degree in mechanical engineering from Beijing Institute of Technology, in 2016, and the M.S. degree in electrical and electronic engineering from Nanyang Technological University, in 2019. She is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong. Her research interests include computer vision, computational imaging and machine learning.

Edmund Y. Lam received the B.S., M.S. and Ph.D. degrees in electrical engineering from Stanford University. He was a Visiting Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, from 2010 to 2011, He is currently a Professor in electrical and electronic engineering with The University of Hong Kong, and serves as the Computer Engineering Program Director. His main research is computational imaging. He is a Fellow of the IEEE, OSA, SPIE, IS&T and HKIE, and was a recipient of the IBM Faculty Award.