Light Field Image Restoration via Latent Diffusion and Multi-View Attention

Shansi Zhang, *Graduate Student Member, IEEE*, and Edmund Y. Lam, *Fellow, IEEE*

***Abstract*—Light field (LF) images contain information for multiple views. The restoration of degraded LF images is of great significance for various LF applications. Inspired by the recent achievement of denoising diffusion models, we propose a LF image restoration method based on latent diffusion (LD). We design a LDUNet with efficient cross-attention modules to integrate the features of conditional input, and propose a two-stage training strategy, where the LDUNet is first trained on the individual views and then fine-tuned on the LF images with injected prior noise. A refinement module is jointly trained in the second stage to enhance the spatial-angular structures. It consists of multi-view attention blocks with patch-based angular self-attention to fuse the global view information. Moreover, we introduce an enhanced noise loss for better noise prediction and an auxiliary image loss to obtain high-quality images. We evaluate our method on LF image deraining task and low-light LF image enhancement task. Our method demonstrates superior performance on both tasks compared to the existing methods.**

***Index Terms*—Cross-attention, latent diffusion, light field, prior noise, multi-view attention.**

I. Introduction

LIGHT field (LF) refers to the collection of light rays in space. A plenoptic camera can capture the entire light field of a scene in a single shot to produce LF images [1], [2]. Multiple sub-aperture images (SAIs) that record the scene from different perspectives can be extracted from each raw LF image. This property enables many applications, such as post-capture refocusing [3], depth estimation [4], [5], scene reconstruction [6], and salient object detection [7]. However, these applications could be affected by various degradations caused by bad weathers, low-light conditions, motion blur, etc. Therefore, LF image restoration algorithms should be developed in order to obtain high-quality LF images for better applications.

Nowadays, deep learning has demonstrated greater power in image restoration tasks compared to traditional methods. Various advanced architectures based on the convolutional neural network (CNN) and vision transformer are designed to achieve superior performance. Different from single images, the research on LF image restoration should focus on not only the feature extraction within the individual views but also the utilization of multi-view information. There have been some methods developed for specific LF image restoration tasks, such as rain removal [8], [9], low-light enhancement [10], [11], [12], [13], [14], [15] and deblurring [16], [17]. Regarding LF image deraining, the existing methods [8] and [9] both involve a complex pipeline with separate networks to achieve depth estimation and rain streak removal. They also adopt the generative adversarial network (GAN) architecture to improve image qualities. Regarding low-light LF enhancement, various frameworks were developed, including the L3Fnet [12] with global representation block to encode LF geometries, deep Retinex frameworks [11], [18] with LF decomposition and enhancement, a three-stage framework [13] with global activation, view selection and RNN-based feedforward network, and the LRT [14] that contains spatial-angular transformer blocks and multiple heads to implement intermediate tasks. Regarding LF image deblurring, a deep recurrent network [16] and a view adaptive network [17] were developed to solve the problem of LF motion blur. Instead of focusing on task-specialized solutions, our objective is to develop a unified architecture that can be effectively applied to various LF restoration tasks.

Diffusion models have recently demonstrated remarkable performance in generating high-quality images. Denoising diffusion probabilistic models [19] (DDPMs) were first proposed, which learn to denoise a sample iteratively from a noise distribution. Afterwards, some variant diffusion models were proposed, including the denoising diffusion implicit models (DDIMs) [20], which remove the need of simulating a Markov chain to accelerate sampling, and the latent diffusion models [21], which are trained on the latent space to reduce computational cost and memory consumption. Diffusion models have also been proven to be effective for image restoration tasks by taking the degraded image as the conditional input. Some corresponding methods were developed to achieve deblurring [22], super-resolution [23], low-light enhancement [24] and multi-weather restoration [25] for single images.

Along this line, we present a LF image restoration method based on latent diffusion (LD). Specifically, we propose a LDUNet with efficient cross-attention modules to perform denoising in the latent space. Our cross-attention modules employ intermediate tokens with small size to integrate the features of input image to the denoising process, which involves less computation than the common operations. To fully utilize the views of LF images, the LDUNet is first trained on the individual views. Then, it is fine-tuned on the LF images with injected prior noise by considering the correlation among the views, which leads to an enhanced noise loss for better noise prediction. In addition, we introduce a refinement module with multi-view

Manuscript received 3 January 2024; revised 24 March 2024; accepted 27 March 2024. Date of publication 1 April 2024; date of current version 19 April 2024. This work was supported in part by the Research Grants Council of Hong Kong under Grant GRF 17201822 and in part by the ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong, SAR. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Liu. *(Corresponding author: Edmund Y. Lam.)*

The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR 999077, China (e-mail: shansi@connect.hku.hk; elam@eee.hku.hk).

Digital Object Identifier 10.1109/LSP.2024.3383798

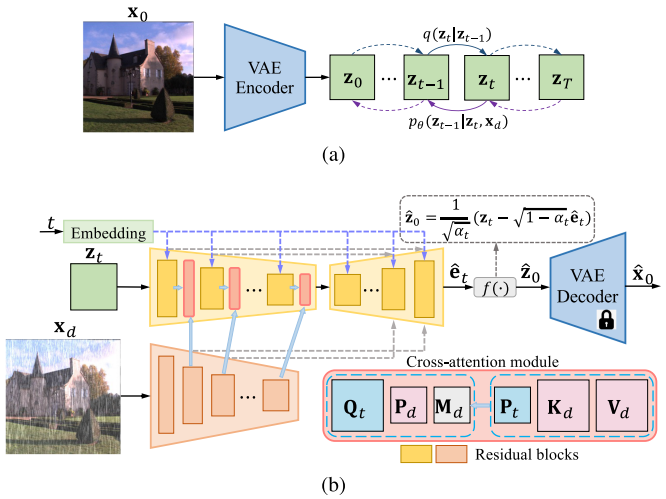

Fig. 1. (a) Diffusion process (blue line) and reverse process (purple line) in latent space. (b) Architecture of LDUNet with efficient cross-attention modules.

attention blocks to fuse the global angular information for better spatial-angular structures, which is jointly trained with the LDUNet in the second stage. Our method demonstrates superior performance on both the LF image deraining and low-light LF image enhancement tasks.

II. Method

A LF image is represented as $\mathbf{X} \in \mathbb{R}^{U\times V\times 3\times H\times W}$, which consists of $U \times V$ SAIs with spatial resolution $H \times W$. A single SAI (view) is denoted as $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$. In what follows, we will introduce the architecture of our diffusion network, LDUNet, the training pipeline for LF image restoration and the multi-view attention in refinement module in detail.

A. LDUNet

Diffusion models [19] are a kind of generative models that learn a data distribution with iterative denoising. However, training diffusion models in the pixel space for high-resolution image generation causes huge computational cost. Thus, we employ the latent diffusion [21] that operates on the low-dimensional latent space to reduce computational complexity and improve efficiency.

We first train a VAE [26] with an encoder $\mathcal{E}$ to encode an image $\mathbf{x}_0 \in \mathbb{R}^{3\times H\times W}$ into a latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0) \in \mathbb{R}^{c_0\times \frac{H}{k}\times \frac{W}{k}}$ (k is a downsampling factor) and a decoder $\mathcal{D}$ to reconstruct the image from $\mathbf{z}_0$. The diffusion process sequentially adds Gaussian noise to $\mathbf{z}_0$, formulated by $q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}\right)$, as shown in Fig. 1(a). We can marginalize the diffusion process to obtain $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{e}_t$ at any time step, where $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, and $\mathbf{e}_t \sim \mathcal{N}(0, \mathbf{I})$.

We design a LDUNet, as depicted in Fig. 1(b), which takes as inputs $\mathbf{z}_t$, t after embedding and a degraded image $\mathbf{x}_d$, and outputs the predicted noise $\hat{\mathbf{e}}_t = \epsilon_\theta(\mathbf{z}_t, \mathbf{x}_d, t)$ injected to $\mathbf{z}_0$. $\mathbf{z}_t$ and $\mathbf{x}_d$ are first fed to separate feature extractors that consist of residual blocks to gradually downsample the features. We propose an efficient cross-attention (ECA) module to integrate the features of $\mathbf{x}_d$ to the features of $\mathbf{z}_t$ at different scales. $\mathbf{Q}_t \in \mathbb{R}^{hw\times c}$ is obtained from the feature of $\mathbf{z}_t$ while $\mathbf{K}_d \in \mathbb{R}^{hw\times c}$ and $\mathbf{V}_d \in \mathbb{R}^{hw\times c}$ are obtained from the feature of $\mathbf{x}_d$ by linear layers. Instead of computing cross-attention directly using $\mathbf{Q}_t$, $\mathbf{K}_d$ and $\mathbf{V}_d$, we introduce two pooled tokens $\mathbf{P}_t \in \mathbb{R}^{\frac{hw}{r^2}\times c}$ and $\mathbf{P}_d \in \mathbb{R}^{\frac{hw}{r^2}\times c}$, derived by average pooling with a linear projection on the corresponding features, where r is a pooling factor. The calculation procedures are as follows

$$\mathbf{M}_d = \text{Softmax}\left(\frac{\mathbf{P}_t\mathbf{K}_d^T}{\sqrt{c}}\right)\mathbf{V}_d \tag{1}$$

$$\mathbf{F}_{\text{CA}} = \text{Softmax}\left(\frac{\mathbf{Q}_t\mathbf{P}_d^T}{\sqrt{c}}\right)\mathbf{M}_d, \tag{2}$$

where $\mathbf{M}_d \in \mathbb{R}^{\frac{hw}{r^2}\times c}$ is the intermediate token and $\mathbf{F}_{\text{CA}} \in \mathbb{R}^{hw\times c}$ is the feature after cross-attention.

Given the feature size $c \times h \times w$, the computational complexity of our cross-attention is calculated by

$$\Omega_{\text{ECA}} = 4hwc^2 + \frac{2hwc^2}{r^2} + \frac{4(hw)^2c}{r^2}, \tag{3}$$

where r is set to 4 and c is in the range $32 \sim 256$. Thus, our cross-attention module involves less computation than the common cross-attention operation with complexity $4hwc^2 + 2(hw)^2c$.

B. Training Pipeline

The training of diffusion models usually requires a large amount of data. However, there is a scarcity of large datasets for LF image restoration. Thus, we first train the LDUNet on the individual views of LF images so that more images can be used for training. The training objective is to minimize the error of noise prediction, expressed as

$$\mathcal{L}_{\text{ns}} = \mathbb{E}_{\mathbf{z}_t,\mathbf{x}_d,\mathbf{e}_t,t}\left[\|\mathbf{e}_t - \epsilon_\theta(\mathbf{z}_t, \mathbf{x}_d, t)\|_1\right]. \tag{4}$$

Moreover, we can obtain the predicted latent representation $\hat{\mathbf{z}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{z}_t, \mathbf{x}_d, t)\right)$. $\hat{\mathbf{z}}_0$ is fed to the VAE decoder to obtain the predicted image $\hat{\mathbf{x}}_0 = \mathcal{D}(\hat{\mathbf{z}}_0)$. To improve the visual quality of generated images, we introduce an image loss term $\ell(\cdot)$ comprising of content loss with ℓ_1 norm, SSIM loss and perceptual loss, expressed as

$$\begin{aligned}\mathcal{L}_{\text{im}} = \ell(\hat{\mathbf{x}}_0, \mathbf{x}_0) = \|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_1 + \text{SSIM}(\hat{\mathbf{x}}_0, \mathbf{x}_0) \\ + \|\phi(\hat{\mathbf{x}}_0) - \phi(\mathbf{x}_0)\|_1,\end{aligned} \tag{5}$$

where $\phi(\cdot)$ denotes extracted deep features from a pretrained VGG network [27]. Thus, the full loss for the first-stage training is $\mathcal{L}_1 = \mathcal{L}_{\text{ns}} + \mathcal{L}_{\text{im}}$.

In the second stage, we further fine-tune the LDUNet on the LF images while enhancing the spatial-angular structures with a refinement module, as shown in Fig. 2. The views in a LF image are interrelated and share similar contents. If the noise added to each view is independent, the diffusion process may ignore the relationship among the views and therefore cause the inconsistency of restored spatial information. Given this, we introduce the prior noise $\mathbf{E}_t \in \mathbb{R}^{U\times V\times c_0\times \frac{H}{k}\times \frac{W}{k}}$ injected to each LF image, where the noise samples for the individual views are correlated. The noise added to the central view is denoted as $\mathbf{e}_t^c \sim \mathcal{N}(0, \mathbf{I})$, and the noise added to each peripheral view is represented as

$$\mathbf{e}_t^p = \sqrt{\gamma}\mathbf{e}_t^c + \sqrt{1-\gamma}\mathbf{e}_t^i, \tag{6}$$

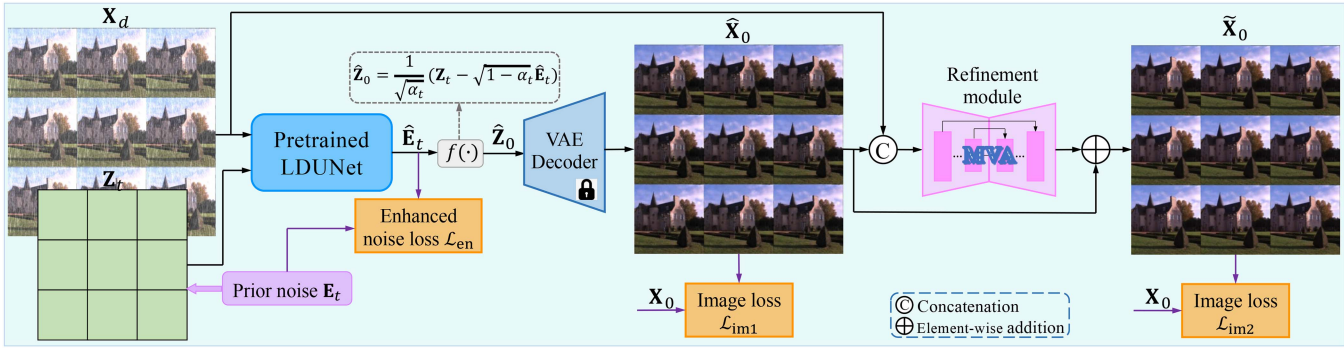

Fig. 2. Training pipeline for LF images. The LDUNet is fine-tuned with injected prior noise, and a refinement module is trained jointly to improve spatial-angular structures.

where $\mathbf{e}_t^i \sim \mathcal{N}(0, \mathbf{I})$ is the individual noise per view and $\gamma \in [0, 1]$ denotes the ratio of the shared noise $\mathbf{e}_t^c$. $\mathbf{E}_t$ consists of $\mathbf{e}_t^c$ for the central view and $\mathbf{e}_t^p$ for all the peripheral views.

During training, we first obtain $\mathbf{Z}_0 = \mathcal{E}(\mathbf{X}_0)$ and $\mathbf{Z}_t = \sqrt{\bar{\alpha}_t}\mathbf{Z}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{E}_t$ for each LF image. Then, $\mathbf{Z}_t$ and degraded LF image $\mathbf{X}_d$ are fed to the pretrained LDUNet to output estimated noise $\hat{\mathbf{E}}_t$, with $\hat{\mathbf{e}}_t^c$ for the central view and $\hat{\mathbf{e}}_t^p$ for the peripheral views. The estimated individual noise can be obtained by $\hat{\mathbf{e}}_t^i = \frac{1}{\sqrt{1-\gamma}}(\hat{\mathbf{e}}_t^p - \sqrt{\gamma}\hat{\mathbf{e}}_t^c)$. Thus, we introduce an enhanced noise loss to provide stronger supervision for noise prediction, with

$$\mathcal{L}_{\text{ens}} = \|\mathbf{E}_t - \hat{\mathbf{E}}_t\|_1 + \|\mathbf{e}_t^c - \hat{\mathbf{e}}_t^c\|_1 + \sum \|\mathbf{e}_t^i - \hat{\mathbf{e}}_t^i\|_1, \quad (7)$$

where $\sum$ denotes summation over all the peripheral views.

Then, we can obtain $\hat{\mathbf{Z}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{Z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\mathbf{E}}_t\right)$, which is fed to the VAE decoder to output $\hat{\mathbf{X}}_0 = \mathcal{D}(\hat{\mathbf{Z}}_0)$. The concatenation of $\hat{\mathbf{X}}_0$ and $\mathbf{X}_d$ is further passed through a refinement module $\mathcal{R}$ with multi-view attention blocks to refine the spatial contents and angular geometries, obtaining the final restored LF image $\tilde{\mathbf{X}}_0$, with

$$\tilde{\mathbf{X}}_0 = \mathcal{R}\left(\left[\hat{\mathbf{X}}_0, \mathbf{X}_d\right]\right) + \hat{\mathbf{X}}_0, \quad (8)$$

where $[\cdot]$ denotes concatenation.

The image loss is applied to both $\hat{\mathbf{X}}_0$ and $\tilde{\mathbf{X}}_0$, with $\mathcal{L}_{\text{im1}} = \ell(\hat{\mathbf{X}}_0, \mathbf{X}_0)$ and $\mathcal{L}_{\text{im2}} = \ell(\tilde{\mathbf{X}}_0, \mathbf{X}_0)$. Thus, the full loss for the second-stage training is written as

$$\mathcal{L}_2 = \mathcal{L}_{\text{ens}} + \mathcal{L}_{\text{im1}} + \mathcal{L}_{\text{im2}}. \quad (9)$$

C. Multi-View Attention

Fig. 3 shows the structure of our multi-view attention (MVA) mechanism, which aims to fuse the global view information for assisting the restoration of each individual view. The input feature $\mathbf{F}_{\text{in}} \in \mathbb{R}^{U\times V\times c\times h\times w}$ is first divided into n groups along the channel dimension, and the feature of each view is divided into $m \times m$ patches. Then, all the patches are fed to a global average pooling (GAP) layer to obtain the patch tokens, which are further processed by the layer normalization (LN) and linear layers with parameter matrices $\mathbf{W}_Q$ and $\mathbf{W}_K$ to derive $\mathbf{Q}_a \in \mathbb{R}^{m^2n\times UV\times \frac{c}{n}}$ and $\mathbf{K}_a \in \mathbb{R}^{m^2n\times UV\times \frac{c}{n}}$. Self-attention is computed using the patch tokens from the same location of each view to obtain

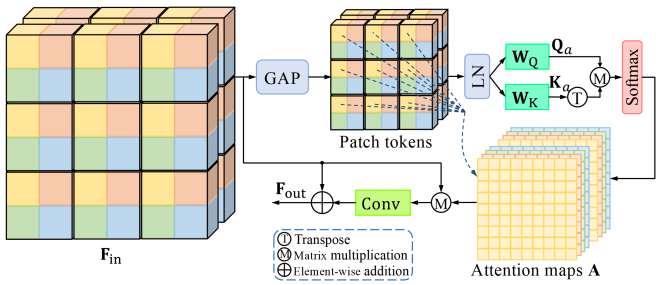

Fig. 3. Multi-view attention with patch-based angular self-attention used in the refinement module.

the attention maps $\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}_a\mathbf{K}_a^T}{\sqrt{c/n}}\right) \in \mathbb{R}^{m^2n\times UV\times UV}$. The multiplication of attention maps and reshaped $\mathbf{F}_{\text{in}}$ is passed through a 3×3 convolution layer, and then added to $\mathbf{F}_{\text{in}}$ to obtain the output feature $\mathbf{F}_{\text{out}} = \text{Conv}(\mathbf{A}\mathbf{F}_{\text{in}}) + \mathbf{F}_{\text{in}}$.

The refinement module consists of several residual blocks for spatial feature extraction and multi-view attention blocks, which operate on different scales and are equipped with skip connections.

III. Experiments

A. Implementation Details

Regarding the VAE, its encoder downsamples the image by a factor $k = 8$ and the channel number of the latent representation is $c_0 = 4$. The VAE was trained on the individual views of LF images with a learning rate 1×10^{-3} for 600 epochs. The LDUNet was trained for 500 epochs with a learning rate 5×10^{-4} in the first stage. The LDUNet and the refinement module were jointed trained for additional 500 epochs with leaning rates 1×10^{-4} and 5×10^{-4}, respectively, in the second stage. Adam optimizer [28] was used in all the training processes. During testing period, we employ the sampling approach in DDIM [20] to accelerate inference.

B. Comparison

We apply our diffusion-based method on two tasks, LF image deraining and low-light LF image enhancement. Regarding LF image deraining, the condition input to the LDUNet is the rainy LF images. We compared our method with two single image deraining methods, BRN [29] and DRT [30], and two LF

TABLE I
QUANTITATIVE RESULTS ON LF IMAGE DERAINING

Method	PSNR↑	SSIM↑	LPIPS↓
BRN [29]	27.480	0.947	0.024
DRT [30]	26.479	0.909	0.048
DEBRNet [8]	26.802	0.903	0.046
MGPDNet [9]	28.654	0.930	0.034
Ours	**32.200**	**0.957**	**0.021**

The bold values are indicates the best results.

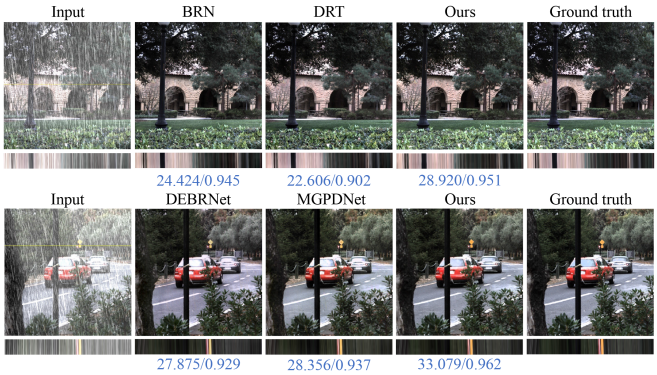

Fig. 4. Visual results of different deraining methods. Zoom in for best view.

TABLE II
QUANTITATIVE RESULTS ON LOW-LIGHT LF IMAGE ENHANCEMENT

Method	Synthetic			Real-world		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
LFRetinex [11]	21.610	0.719	0.182	21.896	0.702	0.175
L3Fnet [12]	21.916	0.737	0.192	22.110	0.722	0.166
TSNet [13]	22.431	0.773	0.171	22.566	0.730	0.161
PMSNet [15]	24.291	0.815	0.125	23.627	0.789	0.121
Ours	**26.152**	**0.832**	**0.100**	**25.866**	**0.823**	**0.094**

The bold values are indicates the best results.

image deraining methods, DEBRNet [8] and MGPDNet [9]. We trained and evaluated all these methods on the dataset provided by [9]. The quantitative results in terms of PSNR, SSIM [31] and LPIPS [32] are listed in Table I. Some visual results including the central views and epipolar-plane images (EPIs) with corresponding PSNR/SSIM score are shown in Fig. 4. We can observe that our method achieves better performance on LF deraining with better quantitative results and visual qualities compared to the other methods.

Regarding low-light LF enhancement, the condition input is the concatenation of low-light LF image and its color map obtained by the approach in [24]. We compared our method with the existing low-light LF enhancement methods, including LFRetinex [11], L3Fnet [12], TSNet [13] and PMSNet [15]. We trained all the methods on the synthetic dark LF images using the synthesis pipeline in [14] and some real-world dark LF images from the L3F dataset [12]. We then evaluated these methods on both the synthetic and real-world dark LF images in the test sets. The quantitative results are listed in Table II and some visual results are presented in Fig. 5. It can be seen that our method outperforms the other methods and obtains higher-quality LF images in terms of illumination, color and object details.

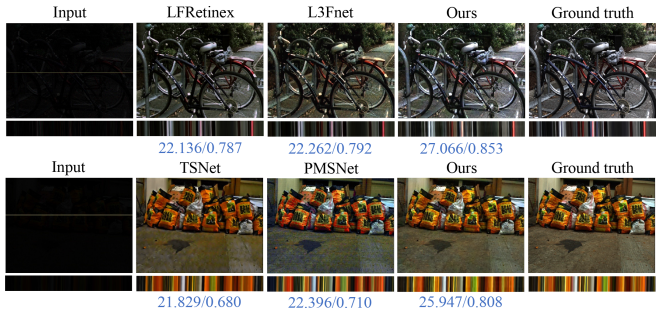

Fig. 5. Visual results of different low-light LF enhancement methods. Zoom in for best view.

TABLE III
ABLATION STUDIES ON LF IMAGE DERAINING

ECA	Prior noise	MVA	PSNR↑	SSIM↑	LPIPS↓
✗	✓	✓	30.441	0.934	0.036
✓	✗	✓	31.833	0.953	0.024
✓	✓	✗	30.045	0.926	0.048
✓	✓	✓	**32.200**	**0.957**	**0.021**

The bold values are indicates the best results.

C. Ablation Studies

We conducted several ablation studies to validate our efficient cross-attention module, prior noise for LF images and multi-view attention mechanism, respectively. We first trained an additional model by replacing our cross-attention modules in LDUNet with the modules in [21], which use the conditional feature map with the lowest resolution to calculate cross-attention. The quantitative results on LF image deraining are listed in Table III. We can see that the performance of this configuration (w/o ECA) is worse than our default setting, verifying the effectiveness of our cross-attention module.

We then trained an additional model without the prior noise while replacing the enhanced noise loss with the common noise loss. The quantitative results of this configuration (w/o prior noise) have some decline compared to our default setting, which demonstrates that the prior noise for LF images with enhanced noise loss helps to improve the performance.

We also trained an additional model by replacing our multi-view attention with the angular convolution [33] in the refinement module (w/o MVA). The angular convolution is a commonly used approach to extract angular features. The corresponding results indicate that our multi-view attention contributes to better performance in LF image restoration.

IV. CONCLUSION

We present a diffusion-based LF image restoration method, which contains a LDUNet with efficient cross-attention modules to perform denoising in the latent space. We propose a two-stage training strategy, where the LDUNet is first trained on the individual views and then fine-tuned on the LF images with injected prior noise. A refinement module with multi-view attention blocks is jointly trained in the second stage to enhance the spatial-angular structures.

REFERENCES

[1] N. Meng, H. K.-H. So, X. Sun, and E. Y. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 873–886, Mar. 2021.

[2] Z. Liang, Y. Wang, L. Wang, J. Yang, and S. Zhou, "Light field image super-resolution with transformers," *IEEE Signal Process. Lett.*, vol. 29, pp. 563–567, 2022.

[3] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective light field refocusing for camera arrays using bokeh rendering and super-resolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, Jan. 2019.

[4] N. Meng, K. Li, J. Liu, and E. Y. Lam, "Light field view synthesis via aperture disparity and warping confidence map," *IEEE Trans. Image Process.*, vol. 30, pp. 3908–3921, 2021.

[5] S. Zhang, N. Meng, and E. Y. Lam, "Unsupervised light field depth estimation via multi-view feature matching with occlusion prediction," *IEEE Trans. Circuits Syst. Video Technol.*, early access, doi: 10.1109/TCSVT.2023.3305978.

[6] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light field," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, pp. 1–12.

[7] A. Wang, "Three-stream cross-modal feature aggregation network for light field salient object detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 46–50, 2021.

[8] Y. Ding, M. Li, T. Yan, F. Zhang, Y. Liu, and R. W. H. Lau, "Rain streak removal from light field images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 467–482, Feb. 2022.

[9] T. Yan, M. Li, B. Li, Y. Yang, and R. W. H. Lau, "Rain removal from light field images with 4D convolution and multi-scale Gaussian process," *IEEE Trans. Image Process.*, vol. 32, pp. 921–936, 2023.

[10] S. Zhang and E. Y. Lam, "Learning to restore light fields under low-light imaging," *Neurocomputing*, vol. 456, pp. 76–87, 2021.

[11] S. Zhang and E. Y. Lam, "An effective decomposition-enhancement method to restore light field images captured in the dark," *Signal Process.*, vol. 189, 2021, Art. no. 108279.

[12] M. Lamba, K. K. Rachavarapu, and K. Mitra, "Harnessing multi-view perspective of light fields for low-light imaging," *IEEE Trans. Image Process.*, vol. 30, pp. 1501–1513, 2021.

[13] M. Lamba and K. Mitra, "Fast and efficient restoration of extremely dark light fields," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1361–1370.

[14] S. Zhang, N. Meng, and E. Y. Lam, "LRT: An efficient low-light restoration transformer for dark light field images," *IEEE Trans. Image Process.*, vol. 32, pp. 4314–4326, 2023.

[15] X. Wang, K. Chen, Z. Wang, and W. Huang, "PMSNet: Parallel multi-scale network for accurate low-light light-field image enhancement," *IEEE Trans. Multimedia*, vol. 26, pp. 2041–2055, 2024.

[16] J. S. Lumentut, T. H. Kim, R. Ramamoorthi, and I. K. Park, "Deep recurrent network for fast and full-resolution light field deblurring," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1788–1792, Dec. 2019.

[17] Z. Shen, S. Zhang, Z. Zhang, Q. Chen, X. Dong, and Y. Lin, "View adaptive light field deblurring networks with depth perception," 2023, *arXiv:2303.06860*.

[18] S. Zhang and E. Y. Lam, "A deep retinex framework for light field restoration under low-light conditions," in *Proc. Int. Conf. Pattern Recognit.*, 2022, pp. 2042–2048.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[20] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Representations*, 2021.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10685 –10685.

[22] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16272–16282.

[23] H. Li et al., "SRDiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[24] Y. Yin, D. Xu, C. Tan, P. Liu, Y. Zhao, and Y. Wei, "CLE diffusion: Controllable light enhancement diffusion model," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 8145–8156.

[25] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10346–10357, Aug. 2023.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[29] D. Ren, W. Shang, P. Zhu, Q. Hu, D. Meng, and W. Zuo, "Single image deraining using bilateral recurrent network," *IEEE Trans. Image Process.*, vol. 29, pp. 6852–6863, 2020.

[30] Y. Liang, S. Anwar, and Y. Liu, "DRT: A lightweight single image deraining recursive transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 588–597.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[33] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, "Light field spatial super-resolution using deep efficient spatial-angular separable convolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2319–2330, May 2019.