Enhancing Image Deblurring through Diffusion in Low-Frequency Latent Space

Chunyu Hu, Ziwei Zhang, *Member, IEEE,*, Xin Wang, *Member, IEEE,* Tianyin Liao, Ziyi Yang, Yufei Sun, Wenwu Zhu, *Fellow, IEEE*

***Abstract*—Diffusion models (DMs) have recently shown remarkable results in image deblurring, a vital task in computer vision. Despite their progress, the existing methods suffer from local inconsistency, leading to unrealistic reconstructed image details. To address this challenge, we observe that high-frequency information in the existing DMs can exaggerate and generate unrealistic image details, leading to local inconsistency. On the other hand, low-frequency information usually preserves the global structure of the image, which is more conducive in diffusion to generate realistically deblurred images. Motivated by this observation, we propose a novel framework Diffusion in Low-Frequency latent space (DiffLF) for image deblurring to enhance DM's generative capability while reducing unnatural details. Specifically, we propose to apply DM in the low-frequency latent space, which is naturally formed by progressively downsampling in a U-Net architecture. Further, we realize the U-Net architecture by proposing low-frequency latent space diffusion (LFD) blocks and effective state space (ESS) blocks. Our method integrates the advantages of DMs and Mamba to preserve long-sequence characteristics while maintaining the deblurring ability. To evaluate our proposed method, we conduct extensive experiments on benchmark image deblurring datasets. Experimental results show that our proposed DiffLF outperforms the state-of-the-art approaches.**

***Index Terms*—Diffusion model, image deblurring, mamba, low-frequency latent space.**

I. Introduction

IMAGE deblurring is a classic task in computer vision, which aims to restore high-quality images from blurred ones. The blur can result from various factors such as lack of focus, rapid target motion, or camera shake [1], [2]. This task is particularly challenging, as in real-world scenarios only the blurred images are available, with little knowledge of the blur source or the clear image. Moreover, the blur kernel is both unknown and complex.

Recently, Diffusion Models (DMs) have exhibited impressive performance in various computer vision tasks [3]–[9], including image deblurring [10]–[12]. In general, DMs consist of the forward process and the reverse process, where the diffusion process gradually increases the noise to the image until the Gaussian noise is satisfied, and the reverse process aims to reconstruct the image through denoising from pure white Gaussian noise [13]. Compared to other generative models, such as generative adversarial network (GAN) [14], [15], DMs have the capacity to generate more accurate images without suffering from training instability or pattern collapse.

Nonetheless, when applied to image deblurring, the existing DM-based methods [10]–[12] still suffer from local inconsistency, i.e., the reconstructed images are unrealistic in local details. The drawback of this challenge is two-folded. On the one hand, the inherent diversity of DM may generate details that do not exist in the original image. On the other hand, during the restoration of blurred images, an excessive focus on local details may introduce artifacts, leading to large distortion.

To address this challenge, we analyze how different frequencies affect the denoising abilities of DMs. Figure 1 shows that high-frequency information is usually associated with image details. Deblurring an image using a DM can exaggerate some unrealistic details, which are also reflected in the high-frequency information of deblurred image. The low-frequency information in DMs embodies the global structure and characteristics of an image, such as global layouts and smooth color. For DMs, the low-frequency information is more suitable for generating high-fidelity images. However, existing DMs entangle the high-frequency information and low-frequency information and cannot utilize the valuable information during diffusion for image deblurring.

Based on the above motivations, we propose a novel DiffLF framework, which performs diffusion denoising in the low-frequency latent space, while utilizing state space models (SSMs) at the high-resolution end to enhance long-range dependency representation. Specifically, the multi-scale downsampling of U-Net naturally forms a low-frequency, compact representation space. We propose low-frequency latent space diffusion (LFD) blocks and adopt them in this space to reduce the risk of amplifying high-frequency noise/artifacts and improve global structural consistency. Furthermore, to compensate for the semantic and long-range dependencies before and after downsampling, we design effective state space (ESS) blocks at the first and last layers of U-Net, efficiently modeling the global long-sequence features by Mamba. ESS blocks build a robust context in the encoding phase and guide detailed reconstruction in the decoding phase. We conduct extensive experiments on benchmark image deblurring datasets including GoPro [17], HIDE [18], and RealBlur [19]. Experimental results show that our proposed DiffLF outperforms the state-of-the-art approaches, including both CNN/Transformer-based methods, diffusion-based methods and recent Mamba-

Chunyu Hu, Tianyin Liao, Ziyi Yang and Yufei Sun are with the College of Software, Nankai University, Tianjin, China. (e-mail: {huchunyu, 1120230329, 2120220718}@mail.nankai.edu.cn; yufei_sun@sina.com)

Ziwei Zhang is with the School of Computer Science and Engineering, Beihang University, Beijing, China. (e-mail: zwzhang@buaa.edu.cn)

Xin Wang and Wenwu Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. (e-mail: {xin_wang,wwzhu}@tsinghua.edu.cn)

Corresponding authors: Ziwei Zhang, Yufei Sun and Xin Wang.

This work was supported in part by the National Natural Science Foundation of China (No. 92570113, 62472018), and Science and Technology Program of Tianjin (No. 24HHXCSS00001).

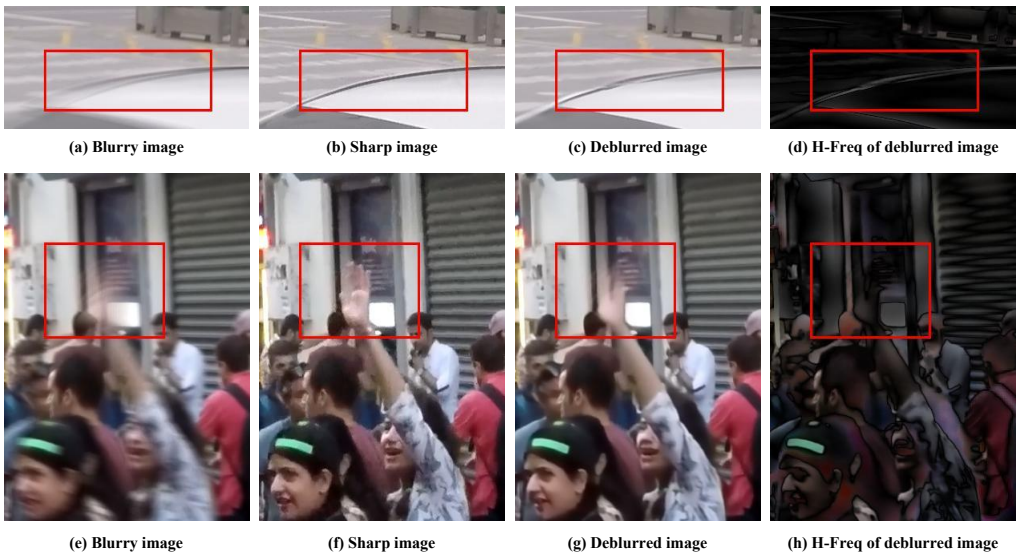

Fig. 1. The deblurring images of an existing representative DM, InDI [16]. The first row corresponds to the outlines of a car and second row corresponds to a human hand. The high-frequency component of the image is extracted using Fourier Transform. It can be observed that high-frequency components are likely to generate or exaggerate some unrealistic details and lead to local inconsistency due to the diffusion on irrelevant artifacts.

based methods. Ablation studies further verify the designs of our proposed method.

The main contributions are summarized as follows:

- We propose a novel framework Diffusion in Low-Frequency latent space (DiffLF), for generating realistic deblurring image while minimizing abnormal details. To the best of our knowledge, DiffLF is the first framework to apply DM in the low-frequency latent space for image deblurring tasks.
- We realize the framework by proposing low-frequency latent space diffusion (LFD) blocks and effective state space (ESS) blocks in a U-Net architecture. Our method combines the advantages of DMs and Mamba by preserving long-sequence characteristics with maintaining the denoising ability.
- We quantitatively and qualitatively evaluate the proposed method on benchmark image deblurring datasets, and show that our proposed approach performs effectively against state-of-the-art methods.

II. Related Work

Image deblurring is a longstanding problem in computer vision with broad applications in surveillance, autonomous driving, and multimedia analysis. It is challenging because multiple seemingly sharp images could lead to the same blurry observation. Some traditional methods tackle this problem by finding an approach that satisfies proximity to particular image and/or blur kernel prior [20]–[22]. With the rapid development of deep learning, many neural network-based methods have been developed. We review three most related categories: CNN/Transformer-based methods, diffusion-based methods, and mamba-based methods.

A. Deep CNN-based and Transformer-based Methods

With the fast development in deep learning, image deblurring techniques have been greatly improved. SRN-DeblurNet [23] propose a multi-scale recurrent network to promote the effective fusion of different scales information. DMPHN [24] put forward a deep hierarchical multi-patch network. To address large blur variations across different spatial locations, SAPHN [25] present an efficient pixel adaptive and feature attentive method. MPRNet [26] propose a multi-stage architecture through feature fusion across stages. MSGSN [27] propose a multi-scale grid network with high-frequency guidance. MDADNet [28] enhances the generalization capability of motion deblurring models in complex real-world scenarios by constructing a multi-domain blur dataset and integrating domain adaptation with meta-deblurring modules. DeepRFT [29] reveals that a simple ReLU operation on the Fourier effectively extracts blur-kernel patterns and serves as a lightweight plug-in block to boost existing networks.

To tackle the above issues, Transformer [30] has been introduced into image deblurring task. Restormer [31] propose an efficient Transformer model for handling high-resolution images. To capture blurred patterns with varying orientations, Stripformer [32] creates intra- and inter-strip tokens to adjust the weighting of image features in both horizontal and vertical directions. Beyond spatial-domain perspective, several works explore image deblurring from a frequency-domain. FFTformer [33] leverages the convolution theorem to convert Transformer self-attention from spatial matrix multiplication to frequency-domain element-wise multiplication, significantly reducing computational cost. LoFormer [34] further proposes a local frequency channel self-attention mechanism to jointly model global low-frequency structures and local high-frequency details. In addition, hybrid CNN-Transformer networks have also been explored. SFHRN [35] leverages a dual-branch attention mechanism to exploit local information in the spatial domain and global information in the frequency domain, respectively, thereby addressing the complex degradation in JPEG compressed images. RUN [36] employs spatial self-attention at high-resolution top levels, channel self-

attention at bottom levels with rich channels, and convolution blocks for mid-level features. Although these methods have achieved commendable performance, their denoising abilities are still limited.

TABLE I
QUALITATIVE COMPARISON WITH PRIOR DIFFUSION AND FREQUENCY-DOMAIN METHODS.

Method	Type	Latent/Freq Space	Diffusion
HI-Diff [37]	LDM	VAE	✓
DiffIR$_{S2}$ [12]	LDM	VAE	✓
DeepRFT [29]	Freq-Domain	Fourier	×
FFTformer [33]	Freq-Domain	Fourier	×
LoFormer-L [34]	Freq-Domain	Fourier	×
SFHRN [35]	Freq-Domain	Fourier	×
DiffLF (Ours)	Low-Freq Latent	U-Net downsampling	✓

B. Diffusion Model-based Methods

DM is a generative model that has recently made significant progress in the fields of image deblurring. DNA [38] enhance conditional diffusion model with a deterministic initial predictor. DiffIR [12] propose a two-stage framework, first using the diffusion model to generate a compact prior representation, then training the DM to estimate the same prior representation from low-quality image. HI-Diff [37] develop the hierarchical integration diffusion model, which introduce DM in a highly compacted latent space for image deblurring. WaveDM [39] accelerates DM-based method by performing the diffusion process in the wavelet domain and employing an efficient conditional sampling strategy. Swintormer [10] introduces a sliding window model for defocus deblurring, leveraging a diffusion model to generate latent prior features. Despite these efforts, these methods still suffer from local inconsistency, resulting in undesired generated artifacts.

To further clarify the distinction between our DiffLF and existing latent diffusion or frequency-domain methods, Table I provides a comparison. Unlike LDMs that operate in a Variational Auto-Encoder(VAE) space, our low-frequency latent space is naturally formed by U-Net downsampling. Unlike frequency-domain methods that explicitly process Fourier without diffusion modal, we perform diffusion in the low-frequency space to enhance global structural consistency while suppressing high-frequency artifacts.

C. Mamba-based Methods

Recently, SSMs [40], [41] have become an effective framework because of its ability to capture long-sequence dependencies with linear complexity. Mamba [42] and its improved version, Mamba-2 [43], have been applied to visual tasks [44], [45], demonstrating strong performance. MambaIR [45] propose an integrated channel attention and Mamba method for image restoration, which involves scanning along four different directions. VmambaIR [44] propose omni selective scan mechanism, which represents the comprehensive scanning and modeling of the input information in six directions. CU-Mamba [46] put forward a channel and spatial state space model block to improve the quality of the restored images.

TABLE II
NOTATIONS USED IN THIS PAPER.

Notation	Description
$\mathbf{I}_{\text{blur}}$	A blurry input image
$\mathbf{R}$	A residual image
$\hat{\mathbf{I}} = \mathbf{I}_{\text{blur}} + \mathbf{R}$	A predicted deblurred image
$\mathbf{F}_s, \mathbf{F}_e, \mathbf{F}_l, \mathbf{F}_d, \mathbf{F}_r$	Image feature maps
$\boldsymbol{x}$, $\boldsymbol{y}$	Sharp and blurry image features in DM
$\boldsymbol{x}_t$	Intermediate degraded image features in DM
$\hat{\boldsymbol{x}}_{t-\delta}$	The predicted deblurred image at time $t-\delta$
$t \in [0, 1]$	The time step
T	The number of steps in DM
$\delta = \frac{1}{T} \in (0, 1)$	A parameter of deblurred speed
$F_\theta(\cdot; t)$	A network of the diffusion process
θ	Parameters of the model
ϵ	A learnable parameter of noise intensity
$\boldsymbol{\zeta}$	A random noise

MS-Mamba [47] introduce a multi-scale SSM-based approach to learn multi-scale representations by global and regional SSM modules. LIEDNet [48] integrates a visual state space module and CNN-based local feature extraction within local Mamba Blocks, achieving efficient restoration by modeling both global and local dependencies. However, these methods are solely based on Mamba for image deblurring, thus unable to take advantage of the powerful capabilities of DM in terms of restoration and generation.

III. METHOD

In this section, we first summarize the main notations used throughout the paper. Notations in this paper are summarized in Table II. Then, we present the overall architecture of the proposed DiffLF and describe the core components: the low-frequency latent space diffusion (LFD) block and the effective state space (ESS) block.

A. Overall architecture

As shown in Figure 2, DiffLF is designed as a three-layer U-Net architecture. Specifically, since the first layer of the encoder and the last layer of the decoder retain the highest image resolution, we adopt SSM-based ESS blocks to handle the long-sequence features. Additionally, in the U-Net architecture, low-frequency information progressively increases while high-frequency information diminishes during downsampling, ultimately forming a latent space predominantly composed of low-frequency features. We apply DM-based LFD blocks to reduce the risk of amplifying high-frequency noise/artifacts and improve global structural consistency.

Mathematically, consider a blurry image $\mathbf{I}_{\text{blur}} \in \mathbb{R}^{H \times W \times 3}$, where $H \times W$ denotes the spatial coordinates. The shallow features $\mathbf{F}_s \in \mathbb{R}^{H \times W \times C}$ is first extracted through a 3×3 convolution layer, where C is the number of channels. Next, these shallow features $\mathbf{F}_s$ are passed through L_1 ESS blocks to generate the feature $\mathbf{F}_e$ that contains the long-range dependencies. Then, through the LFD block, which applies DM in the low-frequency latent space to reconstruct realistic features $\mathbf{F}_l$. Next, we utilize the last decoder formed by the ESS blocks to aggregate the global structure of the DM with the

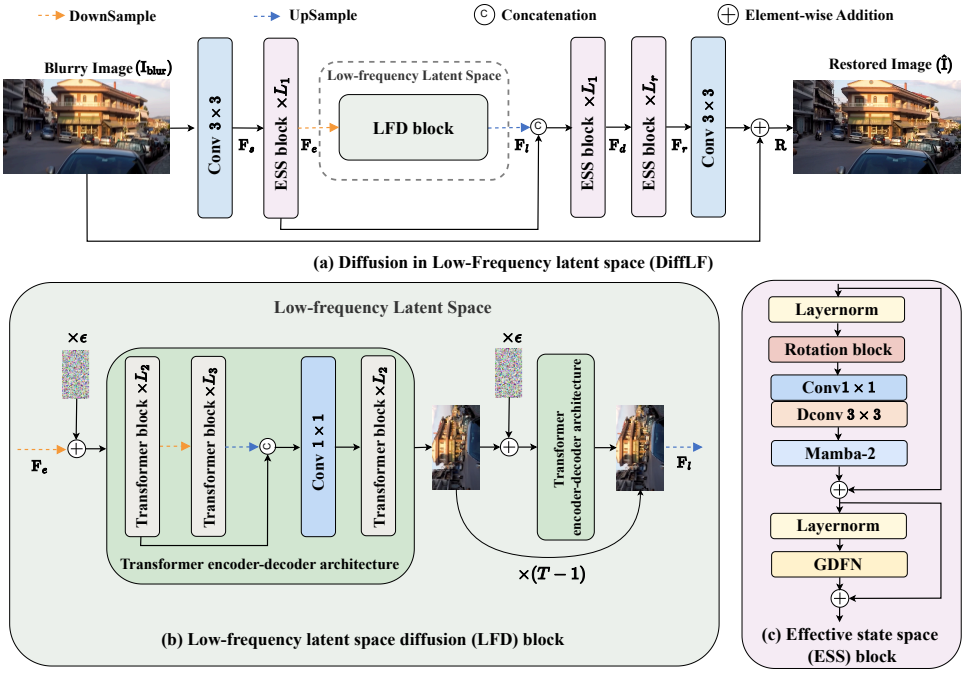

Fig. 2. An overall framework of DiffLF. (a) DiffLF is realized as a U-Net architecture composing of LFD block and ESS blocks for image deblurring. It preserves the long-sequence characteristics of SSMs and the denoising ability of DMs. (b) LFD block applies the diffusion model in the low-frequency latent space. (c) ESS block comprises three primary components: a rotation block, the Mamba-2 to extract long-sequence features, and the Gated-Dconv feed-forward network (GDFN) [31] with a gating mechanism.

high-frequency texture features preserved in the first encoder through skip connection. This ensures that fine-grained details, which are not explicitly processed by the DM, are accurately integrated into the deep feature F_d. To further refine deep features $\mathbf{F}_d$, another set of L_r ESS blocks compose a refinement module to obtain feature $\mathbf{F}_r$. Finally, a 3×3 convolution is applied to the refined feature $\mathbf{F}_r$ to generate the residual image $\mathbf{R} \in \mathbb{R}^{H \times W \times 3}$. The residual image and the blurry image are added together to obtain the reconstructed image: $\hat{\mathbf{I}} = \mathbf{I}_{\text{blur}} + \mathbf{R}$. This residual learning strategy allows the network to focus on modeling the high-frequency degradations and textures while preserving the structural foundations of the input. By learning the difference between the blurry and sharp image rather than reconstructing the entire image, we improve optimization stability and accelerate convergence.The reconstructed image is compared with the ground-truth clear image to calculate loss function and update the model parameters.

B. Latent Space Diffusion Block

In this section, we first define the low-frequency latent space and elaborate on the latent space diffusion block of our method.

The shallow feature $\mathbf{F}_s \in \mathbb{R}^{H \times W \times C}$ are extracted from the blurry input $\mathbf{I}_{\text{blur}}$. Our U-Net encoder applies two downsampling operations, producing features at three scales. After the first downsampling, we obtain $\mathbf{F}_e \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C_1}$ and after the second downsampling we obtain $\mathbf{F}_{e2} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_2}$. We define the low-frequency latent space as the set of features $\{\mathbf{F}_e, \mathbf{F}_{e2}\}$, where progressive downsampling has naturally suppressed high-frequency components.

The existing DMs for image deblurring task [10], [12], [37], [38], [49] suffer from local inconsistency because all spectral information is entangled together. Inspired by [50] and our preliminary observations in Figure 1, high-frequency information is related to image details and when deblurring through the DM, it may result in the generation or exaggeration of unrealistic artifacts. On the other hand, low-frequency components capture the core global elements that define the essence and overall representation of an image. Therefore, we aim to conduct DM in low-frequency latent space to generate more realistic and sharper images. This strategy allows the model to prioritize global structural integrity while effectively suppressing the risk of high-frequency information amplification, ultimately leading to more realistic and perceptually faithful images.

Specifically, we restore blurred images in the latent space. Denote the representation of a sharp image as $\boldsymbol{x}$ and the blurry image as $\boldsymbol{y}$. In our method, the representation is the extracted feature, i.e., $\mathbf{F}_e$. Given $(\boldsymbol{x}, \boldsymbol{y}) \sim p(\boldsymbol{x}, \boldsymbol{y})$, we follow InDI [16], a state-of-the-art DM. The constant forward degradation process is defined as follows:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x} + t\boldsymbol{y} + t\epsilon\boldsymbol{n},\ t \in [0,1], \tag{1}$$

where $\boldsymbol{x}_t$ denotes an intermediate degraded image at time t. The whole degradation process starts from a clear sharp image and gradually adds blur and the Gaussian noise to get a blurry image. ϵ is a parameter to control the strength of noise and $\boldsymbol{n} \sim \mathcal{N}(0, d\mathbf{I})$ is the noise, where $\mathcal{N}$ represents the Gaussian distribution. ϵ is set as a pre-defined constant (*e.g.*, 0.01) in [16]. However, we find that noise should also be moderate

and excessive diversity can result in local inconsistencies in image deblurring. Therefore, we set ϵ as a learnable parameter of the model.

In the reverse process, our goal is to gradually reduce the blur in the image through multiple iterations, eventually obtaining a clear image. Starting from the input degenerate image (when $t = 1$), the optimal reconstruction is generated at time $t - \delta$, where $\delta = \frac{1}{T}$, and T denotes the total number of steps. This optimal reconstruction $\hat{\boldsymbol{x}}_{t-\delta}$ can be estimated by calculating the short-time conditional mean $\hat{\boldsymbol{x}}_{t-\delta} = \mathbb{E}\left[\boldsymbol{x}_{t-\delta} \mid \hat{\boldsymbol{x}}_t\right]$. Though change of variables, the reconstruction process can be derived as InDI [16]:

$$\begin{aligned}\hat{\boldsymbol{x}}_{t-\delta} &= \mathbb{E}\left[\boldsymbol{x}_{t-\delta} \mid \hat{\boldsymbol{x}}_t\right] \\ &= \frac{\delta}{t}\mathbb{E}\left[\boldsymbol{x}_0 \mid \hat{\boldsymbol{x}}_t\right] + \left(1 - \frac{\delta}{t}\right)\hat{\boldsymbol{x}}_t + (t-\delta)\sqrt{\epsilon_{t-\delta}^2 - \epsilon_t^2}\boldsymbol{\zeta},\end{aligned} \tag{2}$$

where a new noise $\boldsymbol{\zeta} \sim \mathcal{N}(0, d\mathbf{I})$ is sampled at each step and incorporated into the current state.

Training DM means training a network $F_\theta(\cdot; t)$ to learn the mapping from intermediate degraded images to original high-quality images. To capture the global information about the image, we construct a network utilizing the Transformer blocks of Restormer [31]. The Transformer blocks operate on the feature dimension rather than the spatial dimension, thereby achieving linear computational complexity. The training goal is to optimize θ by minimizing the expected pixel error

$$\min_\theta \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{t\sim p(t)}\mathbb{E}_{\boldsymbol{n}\sim\mathcal{N}(0,d\mathbf{I})}\left\|F_\theta\left(\boldsymbol{x}_t; t\right) - \boldsymbol{x}\right\|_p, \tag{3}$$

where $p(t)$ represents a predetermined distribution for t. As a result, Eq. (2) can be rewritten as

$$\hat{\boldsymbol{x}}_{t-\delta} = \frac{\delta}{t}F_\theta\left(\hat{\boldsymbol{x}}_t; t\right) + \left(1 - \frac{\delta}{t}\right)\hat{\boldsymbol{x}}_t + (t-\delta)\sqrt{\epsilon_{t-\delta}^2 - \epsilon_t^2}\boldsymbol{\zeta}. \tag{4}$$

During inference, random noise $\boldsymbol{\zeta}$ conforming to the standard normal distribution is generated. Then, by iterating T times according to Eq. (4), a clearer feature $\mathbf{F}_1$ is reconstructed.

C. Effective State Space Block

In this section, we introduce our effective state space (ESS) block, which is a novel modification of Mamba. First, we briefly review the basics of Mamba, and then present our modifications.

Mamba-2 [43] is an improved method of Mamba that applies matrix multiplication based on block decomposition. These models build upon the structured state space sequence models (S4), which is inspired by a particular continuous system. The model maps 1-dimension input sequence $x(t)$ to output sequence $y(t)$ through the hidden state $h(t)$, as follows:

$$\begin{aligned}h'(t) =& \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) =& \mathbf{C}h(t),\end{aligned} \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, and N denotes the state size. The continuous parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ can be transformed into discrete parameters $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, $\overline{\mathbf{C}}$ by employing the zero-order hold (ZOH) with step size Δ:

$$\begin{aligned}\overline{\mathbf{A}} =& \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} =& (\Delta\mathbf{A})^{-1}(\exp(\mathbf{A}) - \mathbf{I}\cdot\Delta\mathbf{B}), \\ \overline{\mathbf{C}} =& \mathbf{C}.\end{aligned} \tag{6}$$

Algorithm 1 The Inference Procedure of DiffLF

Input: Blurry images $\mathbf{I}_{\text{blur}}$, Parameters ϵ, δ, and Trained Blocks in DiffLF

Output: Deblurred images $\hat{\mathbf{I}}$.

```
Calculate F_e = ESS(Conv(I_blur))
/* Reverse Process of DM */
Sample n ~ N(0, dI)
Calculate x_1 = F_e + εn
for t = 1 to 0 with step δ = 1/T do
    Sample ζ ~ N(0, dI)
    Calculate x̂_{t-δ} using Eq. 4
end for
Set F_l = x̂_0
Calculate Î using Eq. 12 and Eq. 13
```

Therefore, the discrete form of models is formulated as:

$$\begin{aligned}\mathbf{h}_t =& \overline{\mathbf{A}}\mathbf{h}_{t-1} + \overline{\mathbf{B}}\mathbf{x}_t, \\ \mathbf{y}_t =& \overline{\mathbf{C}}\mathbf{h}_t.\end{aligned} \tag{7}$$

Based on Eq. 6, selective state space models changes the parameters ($\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, $\overline{\mathbf{C}}$) to be time-varying:

$$\begin{aligned}\mathbf{h}_t =& \mathbf{A}_t\mathbf{h}_{t-1} + \mathbf{B}_t\mathbf{x}_t, \\ \mathbf{y}_t =& \mathbf{C}_t\mathbf{h}_t,\end{aligned} \tag{8}$$

where $\mathbf{x}_t$ denotes the input at time step t, $\mathbf{y}_t$ is the output, and $\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t$ are learnable parameters, The expression $\mathbf{h}_t$ for an arbitrary moment t can be obtained as follows:

$$\mathbf{h}_t = \sum_{s=0}^{t}\mathbf{A}_{t:s}\mathbf{B}_s\mathbf{x}_s, \tag{9}$$

where $\mathbf{A}_{t:s}$ denotes the product of matrices from $\mathbf{A}_t$ to $\mathbf{A}_s$. $\mathbf{y}_t$ can be obtained by multiplying $\mathbf{C}_t$:

$$\mathbf{y}_t = \sum_{s=0}^{t}\mathbf{C}_t\mathbf{A}_{t:s}\mathbf{B}_s\mathbf{x}_s. \tag{10}$$

Vectorizing over the entire time sequence and applying mathematical induction, the matrix transformation form of Eq. 7 can be derived as:

$$\begin{aligned}\mathbf{y} =& \mathrm{SSM}(\mathbf{x}) = \mathbf{M}\mathbf{x}, \\ \mathbf{M}_{ji} :=& \mathbf{C}_j\mathbf{A}_j\cdots\mathbf{A}_{i+1}\mathbf{B}_i,\end{aligned} \tag{11}$$

where $\mathbf{M}$ is a matrix with $\mathbf{M}_{ji}$ representing the mapping relationship from the i-th element of the input vector $\mathbf{x}$ to the j-th element of the output vector $\mathbf{y}$.

Nevertheless, Mamba-2 still suffers from flattening the input, which transforms the image into a 1-dimensional sequence that disregards the inherent spatial position information of the image. In image deblurring task, the restoration of a single pixel usually depends on the neighborhood information of the pixel. Some methods, such as Vision Mamba [51], are proposed to scan simultaneously from four corners of the feature map to retain the spatial information, but they are usually computation intensive and introduce channel redundancy.

To alleviate the above problems, we propose ESS blocks. The design of ESS blocks is motivated by the need to capture global context even for high-frequency information, which

TABLE III
QUANTITATIVE COMPARISONS ON THREE IMAGE DEBLURRING DATASET, INCLUDING GOPRO [17], HIDE [18], AND REALBLUR [19]. THIS TABLE REPORTS DISTORTION-BASED METRICS, INCLUDING PSNR AND SSIM. PSNR IS MEASURED IN DECIBELS (DB). #PARAM. STANDS FOR THE NUMBER OF PARAMETERS (IN MILLIONS). "-" INDICATES THE RESULTS ARE NOT REPORTED IN THE ORIGINAL PAPER. THE BEST RESULTS ARE IN **BOLD**.

Type	Method	Year	GoPro		HIDE		RealBlur-R		RealBlur-J		#Param.
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
GAN Method	DeblurGAN-v2 [52]	CVPR2019	29.55	0.934	26.61	0.875	35.26	0.944	28.70	0.866	60.9
	DBGAN [53]	CVPR2020	31.20	0.940	28.94	0.915	33.78	0.909	-	-	11.6
CNN Method	SRN [23]	CVPR2018	30.26	0.934	28.36	0.904	35.66	0.947	28.56	0.867	6.8
	DMPHN [24]	CVPR2019	31.20	0.945	29.09	0.924	35.70	0.948	28.42	0.860	21.7
	SAPHN [25]	CVPR2020	31.85	0.948	29.98	0.930	-	-	-	-	23.0
	MIMO-Unet+ [54]	ICCV2021	32.45	0.957	29.99	0.930	35.54	0.947	27.63	0.837	16.1
	MPRNet [26]	CVPR2021	32.66	0.959	30.96	0.940	35.99	0.952	28.70	0.87	20.1
	NAFNet [55]	ECCV2022	33.71	0.967	31.31	0.943	35.97	0.951	28.32	0.857	67.9
Transformer Method	Restormer [31]	CVPR2022	32.92	0.961	31.22	0.942	36.19	0.957	28.96	0.879	26.1
	Stripformer [32]	ECCV2022	33.08	0.962	31.03	0.940	36.08	0.954	28.82	0.876	19.7
	Restormer-local [56]	ECCV2022	33.57	0.966	31.49	0.945	36.14	0.956	28.86	0.876	26.1
	FFTformer [33]	CVPR2023	34.21	0.969	31.62	0.946	35.87	0.953	27.75	0.853	16.6
	DDANet [57]	AAAI2023	33.07	0.962	30.64	0.937	35.81	0.951	-	-	16.2
	LoFormer-L [34]	ACM MM2024	34.09	0.969	31.86	0.949	-	-	-	-	49.0
Mamba Method	CU-Mamba [46]	IEEE-MIPR2024	33.53	0.965	31.47	0.944	-	-	-	-	19.7
	XYScanNet [58]	CVPR2025	33.91	0.968	31.74	0.947	-	-	-	-	7.3
DM Method	DNA [38]	CVPR2022	33.23	0.963	30.07	0.928	-	-	-	-	-
	DiffIR$_{S2}$ [12]	ICCV2023	33.20	0.963	31.55	0.947	-	-	-	-	26.9
	HI-Diff [37]	NeurIPS2023	33.33	0.964	31.46	0.945	36.28	**0.958**	29.15	**0.890**	28.5
	Swintormer [10]	ArXiv2024	33.38	0.965	-	-	-	-	-	-	154.9
Ours	**DiffLF**	-	**34.29**	**0.970**	**32.07**	**0.950**	**36.30**	**0.958**	**29.19**	**0.890**	16.5

typically require long-range dependencies to maintain structural coherence. Unlike previous multi-directional scanning methods that incur prohibitive costs, our ESS block achieves this efficiently.Specifically, as depicted in Figure 2(c), we utilize the rotation block to geometrically transform the input without scanning in multiple directions, and then capture and establish the spatial relationship through convolution, thereby preserving the spatial information of the image. This design allows the model to capture multi-orientation dependencies without the prohibitive cost of parallel scanning branches. The input to the ESS block first passes through a normalization layer, followed by a rotation block. The rotation block applies flipping or transformation operations to the input features. After passing through four rotation blocks, the features are effectively scanned at four different orientations, facilitating comprehensive spatial modeling. For odd-numbered blocks, a flip transformation is applied, while even-numbered blocks utilize a transpose transformation. If the total number of ESS blocks is not a multiple of four, an inverse transformation is applied in the end to restore the original spatial structure. After the rotation block, an 1×1 convolution and a depth-wise convolution are added to enhance the spatial features, which further alleviate the shortcomings of flattening. The whole process can be formulated as:

$$\begin{aligned} \mathbf{X}_{\text{RB}} &= \text{RB}(\text{LN}(\mathbf{X}_{\text{input}})), \\ \mathbf{X}_{\text{mamba}} &= \text{Mamba-2}(f^{dwc}_{3\times3}(f^{c}_{1\times1}(\mathbf{X}_{\text{RB}}))) + \mathbf{X}_{\text{RB}}, \end{aligned} \tag{12}$$

where $\text{LN}(\cdot)$ denotes a normalization layer, $\text{RN}(\cdot)$ denotes a rotation block, $f^{c}_{1\times1}(\cdot)$ denotes a convolution with filter size of 1×1 pixel, and $f^{dwc}_{3\times3}(\cdot)$ denotes a depth-wise convolution layer with the filter size of 3×3 pixel.

After the Mamba module, a feed-forward network is usually adopted to further refine the representation. However, uniformly processing all input features cannot screen and enhance critical features. To tackle this issue, we follow Restormer [31] and adopt gated-Dconv feed-forward network (GDFN) to incorporate an gating mechanism and depth-wise convolution. The gating mechanism allows selective information transmission by combining two parallel linear transformation pathways through element-wise multiplication, with one pathway incorporating the GELU non-linear activation [59]. Moreover, depth-wise convolution efficiently encode spatial information, allowing for the extraction and representation of features at a finer granularity. This can be formulated as:

$$\begin{aligned} \mathbf{X}_1, \mathbf{X}_2 &= f^{dwc}_{3\times3}(f^{c}_{1\times1}(\text{LN}(\mathbf{X}_{\text{mamba}}))), \\ \mathbf{X}_{\text{output}} &= f_{1\times1}(\text{GELU}(\mathbf{X}_1 \odot \mathbf{X}_2)), \end{aligned} \tag{13}$$

where $\odot$ denotes the element-wise multiplication and $\text{GELU}(\cdot)$ denotes the GELU activation function [59]. This combination leads to more effective and computationally efficient feature processing.

The inference procedure of our proposed DiffLF is summarized in Algorithm 1.

IV. EXPERIMENTS RESULTS

In this section, we evaluate our method by comparing it with state-of-the-art methods on benchmark datasets. We first introduce the experimental settings. Then, we present the results and some additional analyses.

A. Experimental Settings

1) Datasets: Following the setup in previous works [31], [33], we evaluate our method on three image deblurring bench-

TABLE IV
QUANTITATIVE COMPARISONS ON THREE IMAGE DEBLURRING DATASET, INCLUDING GOPRO [17], HIDE [18], AND REALBLUR [19]. THIS TABLE REPORTS PERCEPTUAL METRICS, INCLUDING LPIPS AND NIQE. #PARAM. STANDS FOR THE NUMBER OF PARAMETERS (IN MILLIONS). "-" INDICATES THE RESULTS ARE NOT REPORTED. THE BEST RESULTS ARE IN **BOLD**.

Type	Method	Year	GoPro		HIDE		RealBlur-R		RealBlur-J		#Param.
			NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	
GAN Method	DeblurGAN-v2 [52]	CVPR2019	**3.68**	0.117	2.96	0.159	-	-	-	-	60.9
	DBGAN [53]	CVPR2020	4.06	0.110	-	-	-	-	-	-	11.6
CNN Method	MIMO-Unet+ [54]	ICCV2021	4.03	0.091	3.24	0.124	-	-	-	-	16.1
	MPRNet [26]	CVPR2021	4.09	0.089	3.46	0.114	-	-	-	-	20.1
	NAFNet [55]	ECCV2022	4.07	0.078	3.22	0.103	5.36	0.064	**3.77**	0.159	67.9
Transformer Method	Restormer [31]	CVPR2022	4.11	0.084	3.41	0.108	5.23	0.058	3.92	0.146	26.1
	Restormer-local [56]	ECCV2022	4.08	0.081	3.42	0.105	-	-	-	-	26.1
	FFTformer [33]	CVPR2023	4.05	**0.071**	3.29	0.096	5.24	0.066	3.90	0.181	16.6
	DDANet [57]	AAAI2023	4.07	0.084	-	-	-	-	-	-	16.2
	LoFormer-L [34]	ACM MM2024	4.05	0.074	3.30	0.093	-	-	-	-	49.0
Mamba Method	XYScanNet [58]	CVPR2025	4.06	0.091	3.41	0.122	-	-	-	-	7.3
DM Method	DNA [38]	CVPR2022	4.07	0.078	**2.93**	**0.092**	-	-	-	-	-
	HI-Diff [37]	NeurIPS2023	4.12	0.080	3.38	0.106	5.32	0.057	3.96	0.136	28.5
Ours	DiffLF	-	4.09	0.073	3.35	0.098	**5.11**	**0.053**	3.85	**0.131**	16.5

marks, including the GoPro dataset [17], HIDE dataset [18], and RealBlur dataset [19]. GoPro contains 2,103 blurry/sharp image pairs for training and 1,111 image pairs for testing. HIDE contains 2,025 image pairs, and RealBlur consists of two sub-datasets, the RealBLur-R and the RealBLur-J dataset. The test sets for both datasets contain 980 image pairs. Note that for HIDE and RealBlur, we follow the protocols of existing methods and directly apply the model trained on the GoPro dataset for testing. The restored images are evaluated using distortion-based metrics, including the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). In addition, we expand our quantitative comparisons to perceptual metrics such as LPIPS [60] and NIQE [61] to evaluate perceptual quality and naturalness. For distortion-based metrics, higher values indicate better reconstruction fidelity, while for perceptual metrics, lower scores correspond to higher perceptual quality and more natural visual appearance.For fair comparison, FLOPs are calculated according to the official implementation, with the input size set to 256×256.

2) Baselines: We compare our method with five groups of baselines, which cover the mainstream state-of-the-art image deblurring methods.

- GAN methods: DeblurGAN-v2 [52], DBGAN [53];
- CNN-based methods: SRN [23], DMPHN [24], SAPHN [25], MIMO-Unet+ [54], MPRNet [26], NAFNet [55] ;
- Transformer-based methods: Restormer [31], Stripformer [32], Restormer-local [56], FFTformer [33], DDANet [57], LoFormer-L [34];
- Mamba methods: CU-Mamba [46], XYScanNet [58];
- DM-based methods: DNA [38], DiffIR$_{S2}$ [12], HI-Diff [37], Swintormer [10].

3) Implementation Details: In our model, the number of ESS blocks is $L_1 = 6$ and the number of Transformer blocks in LFD blocks is $L_2 = 6$ and L_3 = 12. The number of refinement block is $L_r = 4$. The channel number of shallow feature is $C = 48$. In training the DM, the number of steps T is set to 6. We train models with AdamW optimizer (β_1 = 0.9, β_2 = 0.999, weight decay = $1e{-}4$) for 500,000 iterations. The initial value of the learning rate is $2e{-}4$ gradually reduced to $1e{-}6$ with the cosine annealing schedule [62]. Following previous work [12], [31], we adopt a progressive learning and start training with patch size 128×128 and batch size 64. The patch size and batch size pairs are updated to $[(160^2, 40), (256^2, 16)]$ at iterations [120K, 115K, 165K]. It is worth noting that DiffLF is a fully convolutional architecture that is agnostic to input resolution. Our method is implemented with PyTorch and all experiments are performed on NVIDIA Tesla V100 GPUs.

Additional results and anlysis, including failure cases analysis, complexity analysis and generalisability to other degradations are provided in the Appendix.

B. Comparisons with state-of-the-art Methods

To comprehensively evaluate the proposed DiffLF method, we compared it with the state-of-the-art (SOTA) on three public datasets. Quantitative results are measured using distortion-based metrics, such as PSNR and SSIM. In addition, we present qualitative visual comparisons, more visual comparisons in Appendix.

1) Evaluations on the GoPro dataset: The quantitative results are shown in Table III, our method obtains the highest PSNR and SSIM values on the GoPro dataset, showing that it has outstanding deblurring abilities. Compared with the state-of-the-art Transformer-based method, FFTformer [33], our method achieves 0.08 dB PSNR performance improvement while maintaining a similar number of model parameters. Compared with advanced CNN-based methods [55], our approach achieves superior performance with fewer parameters. In addition, compared to the DM-based methods, our model provides a substantial gain of 0.91 dB PSNR. We attribute the clear effectiveness of our proposed method to the advantage of combining SSM methods and DM methods.

Fig. 3. Visual deblurring results on the GoPro dataset [17]. Compared to other methods, our proposed approach achieves the highest PSNR on this image. Additionally, the recovered image exhibits clearer car contours and is almost free of artifacts.

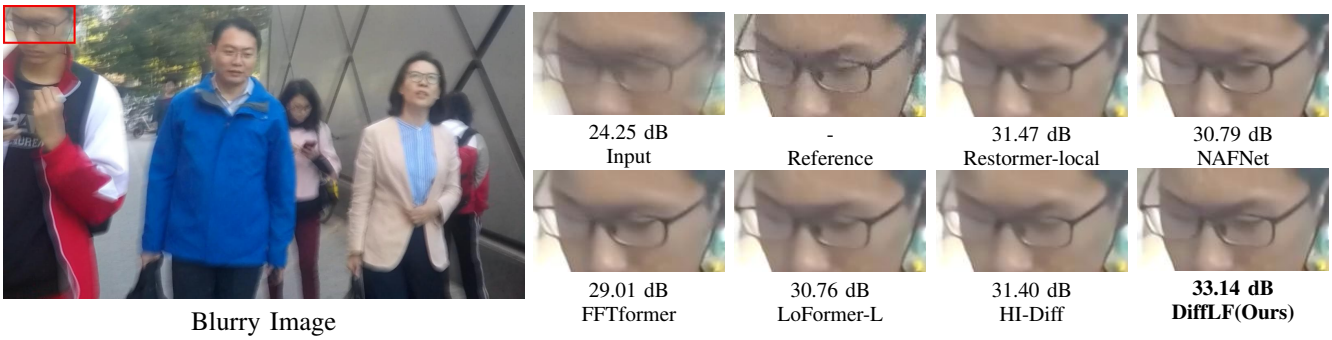

Fig. 4. Visual deblurring results on the HIDE dataset [18]. Other methods are incapable to fully restore the details of the face and glasses well. In contrast, our proposed method recovers images with finer details of the face and glasses.

Beyond distortion-based metrics, we further assess perceptual quality using LPIPS [60] and NIQE [61]. As shown in Table IV, our method achieves the second-best LPIPS score on the GoPro dataset. It is important to note that the perception–distortion tradeoff [63] establishes a fundamental limit, which no image restoration algorithm can simultaneously achieve minimal distortion and perfect perceptual quality. In the perception–distortion plane, there exists an unattainable region. As a method approaches this boundary, any improvement in perceptual quality inevitably leads to increased distortion, and vice versa. Consequently, algorithms optimized for pixel-level losses typically achieve higher fidelity at the expense of perceptual realism, while those emphasizing perceptual metrics generate more visually compelling results with reduced fidelity.

Figure 3 illustrates the visual comparison between the proposed method and baselines on the GoPro dataset. From the comparative images, it can be observed that most competing methods exhibit noticeable artifacts or retain noticeable blur, particularly performing poorly in detail restoration. Taking the DM-based method HI-Diff [37] as an example, it demonstrates unsatisfactory recovery of structural details such as car contours, with noticeably blurred edges. In contrast, our method consistently excels in restoring fine details, producing visual results comparable to the sharp reference images.

2) Evaluations on the HIDE dataset: We further evaluate the generalization capability of our method on the HIDE dataset [18], which contains challenging human-aware motion blur. The quantitative comparisons are shown in Table III. Our method achieves the best performance in terms of both PSNR and SSIM, outperforming all competing approaches. Specifically, our model surpasses the recent Mamba-based method XYScanNet [58] by 0.33 dB in PSNR, while also exceeding DM-based and CNN-based models by a clear margin. These results confirm the strong cross-dataset generalization ability of our method, particularly in handling complex, human-centric blur scenarios that are prevalent in real-world imagery. In terms of perceptual quality, our method also performs favorably on the HIDE dataset [18], achieving competitive LPIPS and NIQE scores as reported in Table IV. This indicates that our approach maintains a good balance between distortion reduction and visual realism, even under more diverse and challenging blur conditions.

The visualization results of Figure 4 show that our proposed method can obtain a relatively high PSNR value. Our approach successfully reconstructs facial regions and eyeglass frames with sharper edges and more authentic textures, whereas other methods fail to recover clearly defined eyeglass structures. The ability to faithfully restore such fine facial features and structural elements underscores the practical value of our method in real-world portrait and video applications.

3) Evaluations on the RealBlur dataset: We also conducted an evaluation of our approach using the RealBlur dataset [19], which is a real-world dataset. In accordance with the protocol adopted by existing methods [31], we directly tested the model trained with the GoPro dataset on the RealBlur dataset. As summarized in Table III, our method achieves competitive performance on both the RealBlur-J and RealBlur-R sub-

TABLE V
ABLATION STUDIES OF THE PROPOSED METHOD ON THE GOPRO DATASET. #PARAM. DENOTES THE NUMBER OF PARAMETERS (MILLIONS).

Method	LFD	ESS	PSNR	SSIM	#Param.
DiffLF (Ours)	✓	✓	33.91	0.9670	7.35
w/o LFD	×	✓	33.25	0.9633	7.30
w/o ESS	✓	×	33.82	0.9669	6.94
w/o ESS & LFD	×	×	32.85	0.9606	6.89

datasets. Notably, our approach attains a PSNR of 36.30 dB on RealBlur-R and 29.19 dB on RealBlur-J, outperforming several recent state-of-the-art methods including Restormer [31] and HI-Diff [37]. This demonstrates the strong cross-domain generalization capability of our method, particularly given that it was trained solely on synthetic blur data from GoPro. The consistent performance gap across different types of real-world blur patterns indicates that our approach learns robust deblurring representations that transfer effectively to practical scenarios. We also present the visualization result in Appendix.

Table IV present perceptual metrics for the RealBlur dataset. On the RealBlur-R dataset, our proposed method achieves the best perceptual quality evaluation results, with an LPIPS value of 0.053 and an NIQE value of 5.11. The LPIPS score is the lowest among all compared methods, indicating that our model performs best in restoring realistic appearance and visual fidelity, with outputs that are perceptually closest to the sharp reference images. On the RealBlur-J dataset, our method also demonstrates strong competitiveness, attaining the best LPIPS score of 0.131. This shows that our approach yields superior visual perceptual quality on JPEG-compressed real-world blurry images. Although our NIQE value is slightly higher than that of NAFNet, when considered together with the LPIPS metric, our method achieves state-of-the-art performance in perceptual similarity.

C. Ablation Studies

In order to validate the effectiveness of out proposed LFD blocks and ESS blocks, we perform ablation experiments on the GoPro dataset. For fair comparisons, the experimental settings are kept the same, with a three-layer U-Net architecture and the number of blocks is [3, 4, 4]. The quantitative results are shown in Table V. We also conducted ablation experiments about the LFD and ESS block in Appendix.

1) Effectiveness of the LFD block: The quantitative results show that LFD blocks are able to bring a considerable improvement in performance to our method, basically having a gain of 1 dB PSNR over the base model (w/o ESS block & LFD block, i.e., the last row). Meanwhile, the model parameters increases only slightly by 0.05M parameters, indicating that the introduction of diffusion modeling in the low-frequency latent space effectively enhances deblurring capability while maintaining parameter efficiency. We argue that the LFD blocks enable the model to focus on reconstructing dominant structural components that reside in the low-frequency information, which stabilizes the optimization process and suppresses overfitting to high-frequency noise. As a result, it helps restore more consistent global details and improves the overall visual fidelity of the deblurred images.

2) Effectiveness of the ESS block: To examine the advantage of our proposed ESS block, we compare it with the base model (w/o ESS block & LFD block). As can be seen, using the ESS block increases the performance from 32.85 dB to 33.25 dB. This demonstrates the effectiveness of the ESS block in capturing long-sequence features and cross-layer contextual information. By modeling the sequential evolution of feature states, ESS block provides richer spatial representations that help to recover fine image textures and local structures.

In short, both the LFD blocks and ESS blocks contribute complementary improvements to our method. The LFD block enhances global consistency and low-frequency reconstruction, while the ESS block refines local structural details. Their combination achieves the highest PSNR and SSIM with only a marginal increase in model parameter, verifying that our method attains an excellent balance between restoration quality and parameter efficiency.

V. CONCLUSION

In this paper, we present a new framework Diffusion in Low-Frequency latent space (DiffLF) for image deblurring, which applies diffusion models in the low-frequency latent space to enhance image deblurring performance. Specifically, we realize the framework by proposing low-frenquency latent space diffusion (LFD) blocks and effective state space (ESS) blocks and combine them in a U-Net architecture, which can leverage the strengths of both SSMs and DMs. Extensive experiments on three benchmark datasets demonstrate that DiffLF outperforms existing state-of-the-art methods. In future work, we aim to extend our method to more image restoration tasks, such as image super-resolution, image denosing, image deraining, image dehazing, and other multimedia applications, such as multimedia content retrieval and social media analysis.

REFERENCES

[1] K. Zhang, W. Ren, W. Luo, W.-S. Lai, B. Stenger, M.-H. Yang, and H. Li, "Deep image deblurring: A survey," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2103–2130, 2022.
[2] C.-H. Liang, Y.-A. Chen, Y.-C. Liu, and W. H. Hsu, "Raw image deblurring," *IEEE Transactions on Multimedia*, vol. 24, pp. 61–72, 2022.
[3] M. Li, T. Cai, J. Cao, Q. Zhang, H. Cai, J. Bai, Y. Jia, K. Li, and S. Han, "Distrifusion: Distributed parallel inference for high-resolution diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7183–7193, 2024.
[4] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, "More control for free! image synthesis with semantic diffusion guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 289–299, 2023.
[5] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.
[6] Z.-C. Zhang, H. Chen, J.-S. Deng, M. Xu, and Z.-B. Pang, "Hexadream: hexaview prior and constraint for text to 3d creation," *Frontiers of Computer Science*, vol. 20, no. 2, pp. 1–14, 2026.
[7] T. Hang, S. Gu, D. Chen, X. Geng, and B. Guo, "Cca: collaborative competitive agents for image editing," *Frontiers of Computer Science*, vol. 19, no. 11, p. 1911367, 2025.
[8] M. Zhou, J. Wang, X. Zhang, D. Campbell, K. Wang, L. Yuan, W. Zhang, and X. Lin, "Probdiffflow: an efficient learning-free framework for probabilistic single-image optical flow estimation," *Frontiers of Computer Science*, vol. 20, no. 8, p. 2008342, 2026.

[9] G. Li, H. Yang, D. Huang, and Y. Wang, "Learning storage-efficient 3d gaussian head avatars from monocular videos via parametric adaptation and material decomposition," *Frontiers of Computer Science*, vol. 20, no. 12, p. 2012711, 2026.

[10] K. Chen and Y. Liu, "Efficient image deblurring networks based on diffusion models," *arXiv preprint arXiv:2401.05907*, 2024.

[11] M. Ren, M. Delbracio, H. Talebi, G. Gerig, and P. Milanfar, "Multiscale structure guided diffusion for image deblurring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10721–10733, 2023.

[12] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13095–13105, 2023.

[13] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, and Z. Chen, "Diffusion models for image restoration and enhancement: a comprehensive survey," *International Journal of Computer Vision*, pp. 1–31, 2025.

[14] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International conference on machine learning*, pp. 537–546, 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[16] M. Delbracio and P. Milanfar, "Inversion by direct iteration: An alternative to denoising diffusion for image restoration," *Transactions on Machine Learning Research*, 2023.

[17] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

[18] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5572–5581, 2019.

[19] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *European Conference on Computer Vision*, pp. 184–201, 2020.

[20] M. Jin, S. Roth, and P. Favaro, "Normalized blind deconvolution," in *Proceedings of the European Conference on Computer Vision*, pp. 668–684, 2018.

[21] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 1964–1971, IEEE, 2009.

[22] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *Acm transactions on graphics*, vol. 27, no. 3, pp. 1–10, 2008.

[23] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8174–8182, 2018.

[24] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5978–5986, 2019.

[25] M. Suin, K. Purohit, and A. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3606–3615, 2020.

[26] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

[27] Y. Liu, F. Fang, T. Wang, J. Li, Y. Sheng, and G. Zhang, "Multi-scale grid network for image deblurring with high-frequency guidance," *IEEE Transactions on Multimedia*, vol. 24, pp. 2890–2901, 2022.

[28] K. Zhuang, Q. Li, Y. Yuan, and Q. Wang, "Multi-domain adaptation for motion deblurring," *IEEE Transactions on Multimedia*, vol. 26, pp. 3676–3688, 2024.

[29] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, "Intriguing findings of frequency selection for image deblurring," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 1905–1913, 2023.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[31] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

[32] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *European conference on computer vision*, pp. 146–162, Springer, 2022.

[33] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5886–5895, 2023.

[34] X. Mao, J. Wang, X. Xie, Q. Li, and Y. Wang, "Loformer: Local frequency transformer for image deblurring," in *ACM Multimedia 2024*, 2024.

[35] S. Tang, H. Zhang, X. Gao, S. Yang, J. Leng, Z. Pan, and H. Tian, "A spatial-frequency hybrid restoration network for jpeg compressed image deblurring," *Neural Networks*, p. 108059, 2025.

[36] Z. Wu, J. Li, C. Xu, D. Huang, and S. C. H. Hoi, "Run: Rethinking the unet architecture for efficient image restoration," *IEEE Transactions on Multimedia*, vol. 26, pp. 10381–10394, 2024.

[37] Z. Chen, Y. Zhang, D. Liu, J. Gu, L. Kong, X. Yuan, *et al.*, "Hierarchical integration diffusion model for realistic image deblurring," *Advances in neural information processing systems*, vol. 36, pp. 29114–29125, 2023.

[38] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16293–16303, 2022.

[39] Y. Huang, J. Huang, J. Liu, M. Yan, Y. Dong, J. Lv, C. Chen, and S. Chen, "Wavedm: Wavelet-based diffusion models for image restoration," *IEEE Transactions on Multimedia*, vol. 26, pp. 7058–7073, 2024.

[40] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

[41] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *In International Conference on Learning Representations*, 2021.

[42] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*.

[43] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," in *International Conference on Machine Learning*, 2024.

[44] Y. Shi, B. Xia, X. Jin, X. Wang, T. Zhao, X. Xia, X. Xiao, and W. Yang, "Vmambair: Visual state space model for image restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[45] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *European Conference on Computer Vision*, pp. 222–241, 2024.

[46] R. Deng and T. Gu, "Cu-mamba: Selective state space models with channel learning for image restoration," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval*, pp. 328–334, IEEE, 2024.

[47] Y. He, L. Peng, Q. Yi, C. Wu, and L. Wang, "Multi-scale representation learning for image restoration with state-space model," *arXiv preprint arXiv:2408.10145*, 2024.

[48] M. Liu, Y. Cui, W. Ren, J. Zhou, and A. C. Knoll, "Liednet: A lightweight network for low-light enhancement and deblurring," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[49] J. Liu, Q. Wang, H. Fan, Y. Wang, Y. Tang, and L. Qu, "Residual denoising diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2773–2783, 2024.

[50] C. Si, Z. Huang, Y. Jiang, and Z. Liu, "Freeu: Free lunch in diffusion u-net," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4733–4743, 2024.

[51] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[52] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8878–8887, 2019.

[53] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li, "Deblurring by realistic blurring," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2737–2746, 2020.

[54] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4641–4650, 2021.

[55] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European conference on computer vision*, pp. 17–33, Springer, 2022.
[56] X. Chu, L. Chen, C. Chen, and X. Lu, "Improving image restoration by revisiting global information aggregation," in *European Conference on Computer Vision*, pp. 53–71, Springer, 2022.
[57] Y. Cui, Y. Tao, W. Ren, and A. Knoll, "Dual-domain attention for image deblurring," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 479–487, 2023.
[58] H. Liu, C. Liu, J. Xu, P. Jiang, and M. Lu, "Xyscannet: A state space model for single image deblurring," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pp. 779–789, June 2025.
[59] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.
[61] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.
[62] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017.
[63] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

Chunyu Hu received the Master's degree from Nankai University, Tianjin, China, in 2022. She is currently pursuing the Ph.D. degree with the College of Software, Nankai University, Tianjin, China. Her research interests include image deblurring, diffusion model and molecular generation.

Ziwei Zhang received his Ph.D. from the Department of Computer Science and Technology, Tsinghua University, in 2021. He is currently an associate professor in the School of Computer Science and Engineering at Beihang University. His research interests focus on machine learning on graphs, including graph neural network (GNN), graph foundation models, and graph-augmented large language models. He has published 50 papers in prestigious conferences and journals, including ICML, NeurIPS, CVPR, AAAI, IEEE TPAMI, and IEEE TKDE.

Xin Wang is currently an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. He got both of his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence, machine learning and its applications. He has published over 200 high-quality research papers in ICML, NeurIPS, IEEE TPAMI, IEEE TKDE, ACM KDD, WWW, ACM SIGIR, ACM Multimedia etc., winning three best paper awards including ACM Multimedia Asia. He is the recipient of ACM China Rising Star Award, IEEE TCMC Rising Star Award and DAMO Academy Young Fellow.

Tianyin Liao received the B.S. degree from Nankai University, Tianjin, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Software, Nankai University, Tianjin, China. His research interests include image deblurring, graph machine learning and out-of-distribution generalization.

Ziyi Yang received the Master's degree from Nankai University, Tianjin, China, in 2025. Her research interest includes computer vision.

Yufei Sun received her Ph.D. from Institute of Computing Technology, Chinese Academy of Sciences, in 2005. She is currently a Professor in the College of Software at Nankai University. Her research interests focus on artificial intelligence, heterogeneous computing, etc. She has published 30 papers in prestigious conferences and journals, including TACO, ASE, ACM KDD, ECCV.

Wenwu Zhu is currently a Professor in the Department of Computer Science and Technology at Tsinghua University. He also serves as the Vice Dean of National Research Center for Information Science and Technology, and the Vice Director of Tsinghua Center for Big Data. Prior to his current post, he was a Senior Researcher and Research Manager at Microsoft Research Asia. He was the Chief Scientist and Director at Intel Research China from 2004 to 2008. He worked at Bell Labs, New Jersey as Member of Technical Staff during 1996-1999. He received his Ph.D. degree from New York University in 1996. His research interests are in the area of data-driven multimedia networking and Crossmedia big data computing. He has published over 400 referred papers and is the inventor or co-inventor of over 100 patents. He received eight Best Paper Awards, including ACM Multimedia 2012 and IEEE Transactions on Circuits and Systems for Video Technology in 2001 and 2019. He served as EiC for IEEE Transactions on Multimedia (2017-2019) and IEEE Transactions on Circuits and Systems for Video Technology (2024- 2025). He served in the steering committee for IEEE Transactions on Multimedia (2015-2016) and IEEE Transactions on Mobile Computing (2007- 2010), respectively. He serves as General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019, respectively. He is an AAAS Fellow, IEEE Fellow, SPIE Fellow, and a member of The Academy of Europe (Academia Europaea).