Hierarchical Wavelet-Guided Diffusion Model for Single Image Deblurring

Xiaopan Li[1], Shiqian Wu[1,2*], Xin Yuan[2,3], Shoulie Xie[4] and Sos Agaian[5]

[1]Institute of Robotics and Intelligent Systems, School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, 430081, China.
[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-Time Industrial System, Wuhan, 430072, China.
[3]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, China.
[4]Institute for Infocomm Research, A*STAR, 138632, Singapore.
[5]Department of Computer Science, College of Staten Island, City University of New York, New York, 10314, USA.

*Corresponding author(s). E-mail(s): shiqian.wu@wust.edu.cn;
Contributing authors: lxp2017@wust.edu.cn; xinyuan@wust.edu.cn; slxie@i2r.a-star.edu.sg; sos.agaian@csi.cuny.edu;

Abstract

Image deblurring is a critical pre-processing step in various high-level vision tasks, including facial recognition, medical imaging, and object detection. Blurred artifacts can arise from multiple factors, such as camera shakes and fast-moving objects. Recently, diffusion models (DMs) have made significant progress in image deblurring by employing a sequence of denoising refinements conditioned on the blurry input. However, existing DM-based methods often neglect the potential of incorporating frequency information, limiting their ability to reconstruct fine textures crucial for high perceptual quality. To this end, we propose a hierarchical wavelet-guided diffusion model (HWDM) for single image deblurring. HWDM integrates multi-level frequency information from wavelet-transformed domains into the denoising network of DM, facilitating the restoration of high-quality deblurred images. Specifically, HWDM consists of three components: primary frequency recovery network (PFRNet), which aims to restore essential frequency information missing in the blurry image; multi-frequency extractor (MFE), which extracts multi-frequency information at various scales from PFRNet's output using multi-level wavelet transforms; and DM's denoising network (DMDNet), into which the frequency information is hierarchically integrated via a cross-attention mechanism, effectively utilizing fine-grained and multi-scale information. Extensive experiments on synthetic and real-world blur datasets demonstrate that HWDM outperforms state-of-the-art methods in perceptual quality, producing more realistic and visually appealing deblurred images. This technology has enormous potential for applications in road traffic, medical imaging, remote sensing satellites, and more. The code is available at https://github.com/TenMiss/HWDiff.

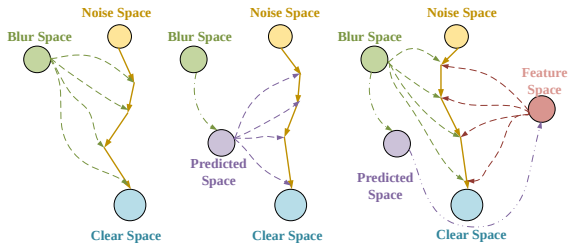
Keywords: Image deblurring, diffusion model, wavelet transform, hierarchical integration

1 Introduction

Single image deblurring aims to restore a clear image from the blurred observation, which is known as an ill-posed inverse problem [1, 2]. Variational optimization was often adopted to solve this issue by considering image and/or kernel priors. Common priors include local maximum gradient prior [3], extreme intensity prior [4], and weighted channel prior [5]. However, these handcrafted priors are insufficient for efficiently modeling the complex and non-uniform blurs in real-world situations.

The emergence of deep learning technologies has revolutionized image deblurring. It is well-known that CNNs are able to learn the implicit priors from vast datasets [6–8]. However, CNNs often learn local features and exhibit translation equivariance properties. In contrast, transformer-based deep neural networks are employed to capture long-range dependencies as priors [9–11]. In recent years, Generative Adversarial Networks (GANs), initially developed for image synthesis, have been successfully applied in the field of image deblurring [12, 13]. GAN-based methods leverage adversarial loss during the training of deblurring networks, resulting in restored images with human perception [14]. However, GANs are prone to several issues, such as model collapse, vanishing gradients, and training difficulties [15].

Recent advances in diffusion models (DMs) [16] have shown impressive capabilities in a variety of computer vision tasks, such as image synthesis [17], image retrieval [18, 19], image super-resolution [20], and image deblurring [21, 22]. DMs leverage denoising autoencoders in the iterative reversal of the diffusion process to produce high-quality mappings from Gaussian noise to target images or latent distributions, effectively avoiding the issues encountered in GANs [23]. In DM-based image deblurring [24], the blurred image guides the reverse process of DM to generate a clear image (Fig. 1 (a)). Recent studies [20, 25] have refined the conditioning process to further improve the quality of the deblurred images by incorporating the predicted images into the DM (Fig. 1 (b)). Consequently, the final quality of the deblurred images is heavily dependent on the predicted images, which may lead to information loss in complex real-world scenarios. A promising strategy leverages both the blurred inputs and robust features as guidance conditions (Fig. 1 (c)). In [22], the guidance was the combination of the blurred image and learned multi-scale structural features extracted through down-sampling



(a) Condition on blurry image (b) Condition on predicted image (c) Condition on predicted feature and blurry image

Fig. 1: Comparison of different conditions in DMs. Solid lines represent the sampling process, dashed lines indicate input conditions, and dash-dotted lines represent the preprocess. (a) Condition: solely the blurred image. (b) Condition: only the predicted image. (c) Conditions: include both the blurred image and the features extracted from the predicted image.

operations, aiming to retain spatial domain information. However, this approach may struggle to distinguish between high-frequency and low-frequency components, potentially limiting its ability to preserve fine details and edge information. In contrast, Discrete Wavelet Transform (DWT) [26–28] excels in multiscale analysis by capturing multi-frequency information across various scales. DWT offers precise time-frequency localization [26] and efficient content-driven feature separation [27], distinguishing it from other frequency extraction techniques such as Discrete Fourier Transform (DFT) [29] and Discrete Cosine Transform (DCT) [30]. Leveraging these advantages, this paper adopts wavelet features to guide the DM, thereby enhancing its capability in restoring fine image details.

In this paper, we present a novel framework for single image deblurring that employs multi-level wavelet features to guide the reverse process of the DM. Initially, PFRNet, a convolutional neural network, generates an initial estimate that preserves more frequency information than the original burred image (see Fig. 2). The loss function of PFRNet is defined in the wavelet domain to further enhance frequency restoration. Subsequently, multi-frequency extractors extract frequency features from PFRNet's output using multi-level wavelet transforms, enhancing its representational capacity across multiple scales and frequencies. These frequency features from wavelet-transformed domains are then hierarchically integrated into the intermediate features of the DM's denoising network (DMDNet) at different scales through a cross-attention module, improving

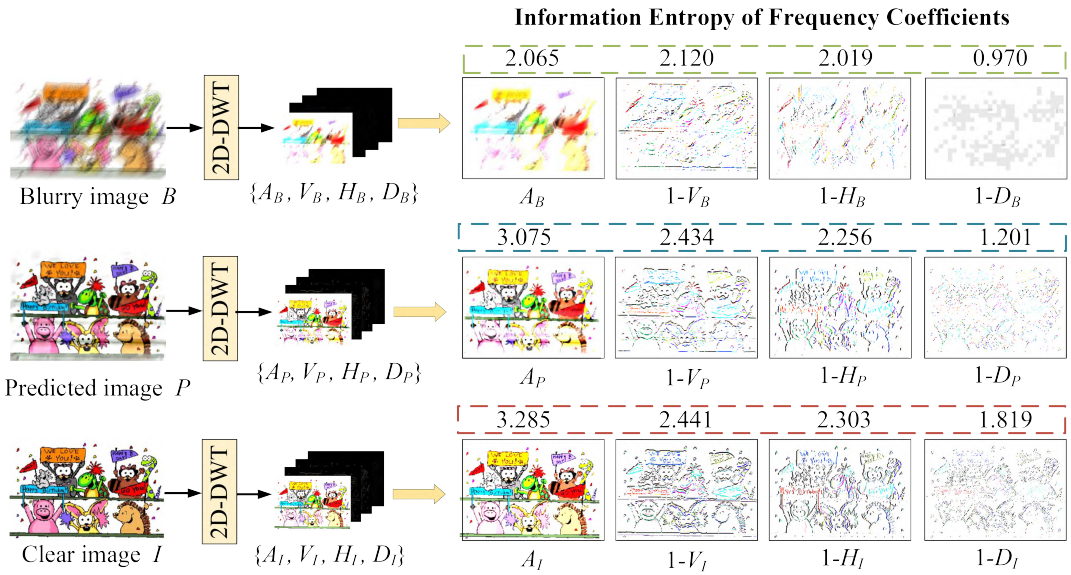


Fig. 2: Illustration of wavelet coefficients for blurry, predicted, and clear images. It is noteworthy that intensity reversal is used for high-frequency bands to facilitate their viewing. Entropy measurement results demonstrate that the predicted image preserves some essential frequency information.

the restoration of clean images in complex blurring scenarios. In addition, we propose a joint optimization of PFRNet and DMDNet to enhance the robustness against estimation errors.

In summary, the main contributions of this work are summarized as follows:

- We propose a novel hierarchical wavelet-guided diffusion model for image deblurring, which integrates multi-level frequency features from wavelet-transformed domains into the reverse process of the diffusion model, enhancing the quality of image restoration.
- We design a primary frequency recovery network to generate an initial estimate with more frequency information than the blurry image, and defined a loss function in the wavelet domain to further enhance frequency recovery capability.
- We introduce a multi-frequency extractor to derive frequency features from PFRNet's output, which are then hierarchically integrated into DM's denoising network using a cross-attention mechanism, enabling comprehensive recovery across various frequency levels.

2 Related Work

2.1 CNN-based Image Deblurring.

With the advent of deep learning, significant progress has been made in image deblurring using CNN-based methods [7, 31–33]. Within the realm of convolutional designs, multi-scale architectures are effective for image deblurring, owing to the complementary information provided by varying scales of the input image. Nah *et al.* [31] introduced a multi-scale deblurring network designed to eliminate motion blur from individual images. Tao *et al.* [32] presented a scale-recurrent network that operates more efficiently across multiple scales for image deblurring. Additionally, Zhang *et al.* [7] developed a deep multi-patch hierarchical network that captures deblurring information from various scales. Recent studies [6, 34] introduced variations to the multi-scale architecture framework through the integration of spatial and channel attentions. This modification aims to hone on crucial information, thereby enhancing overall performance. In this paper, we develop a multi-scale CNN based on multi-level wavelet transforms to generate an initial estimate that contains critical frequency information.

2.2 Transformer-based Image Deblurring.

The vision transformers [35] decompose an image into a sequence of patches (local windows) and learn their mutual relationships by self-attention mechanism. And a lot of transformer-based works have achieved remarkable performance in image deblurring. Zamir *et al.* [11] developed an efficient transformer model that calculates scaled dot-product attention in the feature depth domain, enabling the effective extraction of information across different features along the channel dimension, rather than the spatial dimension. Tsai *et al.* [9] devised horizontal and vertical strip tokens to identify region-specific blurred patterns of varying orientations in dynamic scenes. Moreover, Wang *et al.* [36] introduced a U-shape transformer that employs non-overlapping window-based self-attention for the purpose of single image deblurring.

2.3 Diffusion-based Image Deblurring

The emergence of DMs has marked a significant evolution in the field of generative models. This advancement simplifies the complex and unstable generation process into a sequence of independent and stable reverse processes through Markov Chain modeling. Recently, DMs have yielded encouraging performance in the image deblurring task [21, 22, 37, 38]. Ozan *et al.* [39] introduced a patch-based DM designed for image restoration in adverse weather conditions, which utilizes a guided denoising process that operates across overlapping patches during inference. Some works [37, 38] utilized the DMs to generate a prior representation for image deblurring, and applied a two-stage approach for training. Other approaches [20, 21] integrated DMs with a residual model for image deblurring, achieving competitive values in perceptual-based evaluations. Nonetheless, these methods often suffer from inconsistent detail distribution due to the lack of intermediate constraints, leading to poor performance on distortion-based metrics. Ren *et al.* [22] also highlighted that the absence of intermediate constraints can weaken generalization performance.

2.4 Wavelet-based Image Deblurring

Recent advancements in wavelet-based image deblurring have significantly improved the restoration quality of images. Min *et al.* [40] utilized wavelet transforms to separate frequency information from blurred images, improving recovery by mitigating inherent smoothing effects. Liu *et al.* [26] introduced a multi-layer wavelet transform CNN within a U-net architecture, which enabled adaptive feature adjustment and yielded clearer recovered images. Zou *et al.* [41] proposed a wavelet reconstruction module that effectively helps the network to recover clear high-frequency details. Furthermore, recent studies [14, 20, 27] have integrated wavelets into DMs for image restoration. Shang *et al.* [20] incorporated high-frequency components of DWT while overlooking low-frequency contents. In [14, 27], low-frequency bands were predicted by reverse diffusion processes and high-frequency bands were reconstructed using convolution operations. These methods process different frequency information separately, which limits their ability to recover all frequency signals through the reverse diffusion process. To address this issue, we introduce multi-frequency information as intermediate constraints of the DM through a wavelet-based attention mechanism. This approach informs the DM about the frequency information at intermediate layers, facilitating more effective and robust image deblurring.

3 Proposed Method

In this paper, we develop a new image debluring method by leveraging advantages of both multi-level wavelet transform and diffusion model. Fig. 3 shows the overview of the proposed image deblurring method. Our proposed model is based on the framework of the denoising diffusion implicit model (DDIM) [42], which involves a stochastic diffusion process and a deterministic denoising process.

3.1 Stochastic Diffusion Process

Given a clean sample distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, and gradually adding Gaussian noise according to a variance schedule $\{\beta_1, \beta_2, \cdots, \beta_T\}$, $\mathbf{x}_T$ is close to pure Gaussian noise as T becomes sufficiently large. The forward diffusion process is represented by a parameterized Markov chain:

$$\begin{aligned} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) &= \mathcal{N}\left(\mathbf{x}_t;\ \sqrt{1-\beta_t} \cdot \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \\ q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) &= \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) \end{aligned} \tag{1}$$

where $t = 1, 2, \cdots, T$; $\beta_t \in (0, 1)$ is a predetermined hyperparameter controlling the variance of the noise;

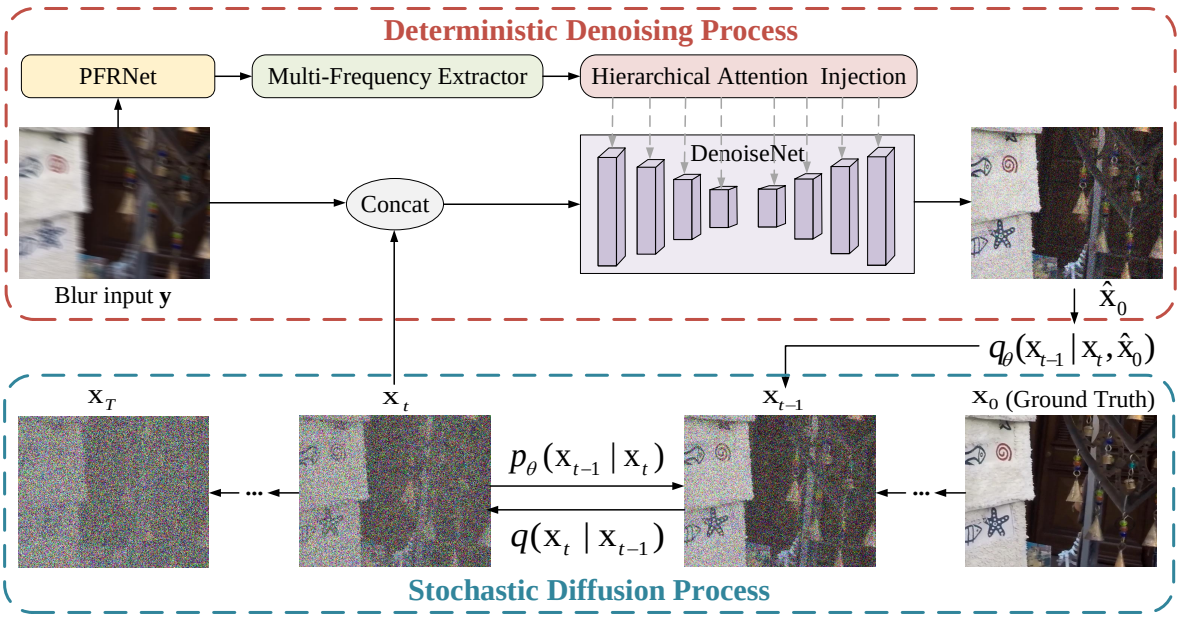


Fig. 3: Illustration of the proposed method. The model involves a stochastic forward diffusion process which gradually adds noise to the clear image (Ground Truth). And a deterministic denoising process is applied to recover clear images corresponding to blurry images. The blurry image and predicted multi-frequency information serve as conditions to guide the denoising process.

$\mathbf{x}_t$ is the noisy image at the t-th step; and $\mathcal{N}$ denotes the Gaussian distribution.

Through iterative derivation with reparameterization, probabilistic distribution of $\mathbf{x}_t$ from $\mathbf{x}_0$ can be computed by

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_t;\ \sqrt{\overline{\alpha}_t}\mathbf{x}_0,\left(1-\overline{\alpha}_t\right)\mathbf{I}\right) \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

The reverse diffusion process of DM reconstructs a sharp result $\mathbf{x}_0$ from a pure Gaussian distribution. By leveraging the editing and data synthesis capabilities of conditional diffusion models [43], conditional denoising process without altering the forward diffusion results in high fidelity of the samples to the distribution conditioned on the conditional input $\mathbf{c}$. Consequently, the distribution of $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ is given by

$$\begin{aligned} q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) &= \mathcal{N}\left(\mathbf{x}_{t-1};\ \mu_t\left(\mathbf{x}_t, t\right), \sigma_t^2 \mathbf{I}\right) \\ \mu_t(\mathbf{x}_t, t) &= \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \cdot \epsilon_t) \end{aligned} \quad (3)$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents the noise in $\mathbf{x}_t$, and is the only uncertain variable. Thus, a neural network $\epsilon_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right)$ estimates the noise ϵ_t for each step.

To optimize the model, we choose to maximize the variational lower bound (VLB) [16]:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{t, \mathbf{x}_0, \epsilon_t}\left[\left\|\epsilon_t - \epsilon_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right)\right\|_2^2\right] \\ &= \mathbb{E}_{t, \mathbf{x}_0, \epsilon_t}\left[\left\|\epsilon_t - \epsilon_\theta\left(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t} \cdot \epsilon_t, \mathbf{c}, t\right)\right\|_2^2\right] \end{aligned} \quad (4)$$

3.2 Deterministic Denoising Process

The reverse process of DM transforms noisy data into clean data samples, which is a stochastic process. DDIMs [42] employ the deterministic generative process $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. Given the image $\mathbf{x}_t$ at step t, the generation process of $\mathbf{x}_{t-1}$ via a forward posterior $q_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \hat{\mathbf{x}}_0)$ can be formulated as follows:

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) &= q_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \hat{\mathbf{x}}_0) \\ &= \mathcal{N}\left(\mathbf{x}_{t-1};\ \sqrt{\overline{\alpha}_{t-1}}\hat{\mathbf{x}}_0 + \sqrt{1-\overline{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t), 0\right) \\ \hat{\mathbf{x}}_0 &= \frac{1}{\sqrt{\overline{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\overline{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)) \end{aligned} \quad (5)$$

To generate realistic deblurred images, we integrate the blurry image $\mathbf{y}$ and the predicted multi-frequency information $\mathbf{z}$, collectively referred to as

Algorithm 1 The inference process of our proposed algorithm

Input: Blurry image $\mathbf{y}$, PFRNet model ϕ_θ, noise predictor ϵ_θ, the time step T, and the number of implicit sampling step S.
Output: Restored image $\mathbf{x}_0$.

1: $\mathbf{z}$ = 2D-DWT $(\phi_\theta(\mathbf{y}))$
2: $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
3: **for** $i = S : 1$ **do**
4: $t = (i-1) \cdot T/(S-1) + 1$
5: $t_{next} = (i-2) \cdot T/(S-1) + 1$ **if** $i > 1$, **else** 0
6: $\mathbf{x}_{t_{next}} = \sqrt{\overline{\alpha}_{t_{next}}} \left(\frac{\mathbf{x}_t - \sqrt{1-\overline{\alpha}_t} \cdot \epsilon_\theta(\mathbf{x}_t, \mathbf{y}, \mathbf{z}, t)}{\sqrt{\overline{\alpha}_t}} \right) + \sqrt{1 - \overline{\alpha}_{t_{next}}} \cdot \epsilon_\theta(\mathbf{x}_t, \mathbf{y}, \mathbf{z}, t)$
7: **end for**
8: **return** $\mathbf{x}_0$

$\mathbf{c}$, into the noise prediction network $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$. The network input is the concatenation of $\mathbf{y}$ and $\mathbf{x}_t$. The multi-frequency information $\mathbf{z}$ is seamlessly integrated into the latent space of each layer via a cross-attention mechanism. Specifically, given a blurry image $\mathbf{y} \in \mathbb{R}^{H \times W \times 3}$, we use PFRNet ϕ_θ to generate a coarse initial estimate, denoted as $\mathbf{p} = \phi_\theta(\mathbf{y})$. Following this, we employ multi-level wavelet transforms to extract multi-frequency coefficients $\mathbf{z}_i$ at multiple scales from the initial estimate. Finally, we introduce a cross-attention mechanism to fuse multi-frequency information into the latent space of the denoising network across multiple scales. This strategy significantly enhances capabilities of the proposed method, particularly in complex blurry scenarios. Algorithm 1 shows the inference process of our method.

3.2.1 Primary Frequency Recovery Network

To capture frequency information, we employ a simple CNN, referred to as PFRNet, to generate an initial prediction of the target image from the blurry image. As shown in Fig. 4, the proposed PFRNet utilizes three levels to extract features, striking a balance between deblurring quality and computational complexity. Here, the two-dimensional discrete wavelet transform (2D-DWT) serves as the downsampling operator, while the two-dimensional inverse wavelet transform (2D-IWT) serves as the upsampling operator. At the bottleneck, three Residual blocks refine the feature representation.

2D-DWT decomposes an image into four sub-bands: A, V, H, and D. This process is reversible, allowing feature extraction without loss of detail information. Among these sub-bands, A preserves the low-frequency content of the original image, while the remaining three contain the high-frequency components of the image from vertical, horizontal, and diagonal directions, respectively. The low-frequency content can undergo further decomposition to obtain multi-scale frequency components. We perform K iterations of wavelet transform on the low-frequency coefficient $A_1 \in \mathrm{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, i.e

$$\{A_{k+1}, V_{k+1}, H_{k+1}, D_{k+1}\} = \text{2D-DWT}(A_k) \quad (6)$$

where A_{k+1}, V_{k+1}, H_{k+1}, $D_{k+1} \in \mathrm{R}^{\frac{H}{2^{k+1}} \times \frac{W}{2^{k+1}} \times C}$, $k \in [1, K]$. It is evident that the spatial dimensions of the wavelet components after wavelet transform are four times smaller than the original image. Therefore, it can be interpreted as the downsampling operator on the image, and is suitable to be applied in multi-scale networks.

3.2.2 Multi-Frequency Extractor

After obtaining the predicted image $\mathbf{p} \in \mathbb{R}^{H \times W \times 3}$ from the PFRNet, we employ 2D-DWT to extract multi-frequency coefficients $\{\hat{A}_i, \hat{V}_i, \hat{H}_i, \hat{D}_i\}$ at the i-th level, i.e.,

$$\{\hat{A}_i, \hat{V}_i, \hat{H}_i, \hat{D}_i\} = \text{2D-DWT}(P) \quad (7)$$

where $\hat{A}_i, \hat{V}_i, \hat{H}_i, \hat{D}_i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times 3}$ for $i = \{1, 2, 3\}$. Executing 2D-DWT three times aims to align with the subsequent designed denoising network of the diffusion model.

Then the proposed multi-frequency extractor (MFE) connects these coefficients by channel to form prior features $\mathbf{z}_i$,

$$\mathbf{z}_i = [\hat{A}_i, \hat{V}_i, \hat{H}_i, \hat{D}_i]_c \quad (8)$$

where $[\cdot]_c$ denotes channel-wise concatenation and $\mathbf{z}_i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times 3 \cdot 4^i}$. The extracted frequency features $\mathbf{z}_i$ are used to guide the diffusion model.

3.2.3 Wavelet-based Attention Module

In typical guided diffusion models for image deblurring, the guidance is the blurred image. However, in our proposed guided diffusion model, the guidance contains not only the blurred image but also the multi-frequency information from the PFRNet. Here

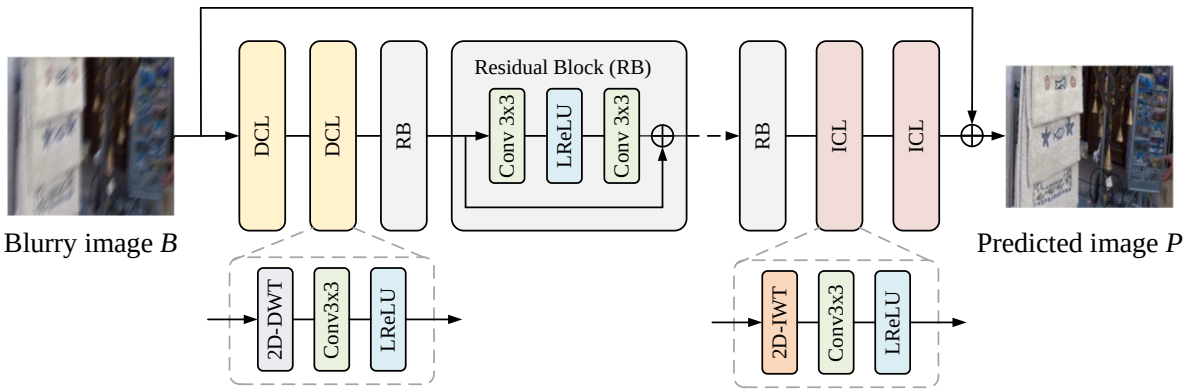


Fig. 4: Network architecture of the proposed PFRNet, where multi-level wavelet transforms are used as downsampling and upsampling operators.

wavelet-based attention module (WAM) is developed to effectively integrate the multi-frequency information into the intermediate transformer-based denoising network of DM (DMDNet), illustrated in Fig. 5 (b). For each WAM, we designed a cross-attention mechanism that employs the multi-frequency feature as the query and the intermediate layer feature as both the key and value. This design enables the model to prioritize wavelet-based multi-scale features, which are essential for capturing high-frequency details in the deblurring process. By using the wavelet-derived multi-frequency feature as the query, the attention mechanism effectively concentrates on these important frequency components within the intermediate layer feature. This approach allows the wavelet-based information to refine intermediate layer feature representation, thereby enhancing the overall performance of the model.

Specifically, through a linear projection on the multi-frequency feature $\mathbf{z}_i \in \mathbb{R}^{h\times w\times c_1}$, the feature map Q containing aggregated frequency information is given by

$$Q = Conv_{1\times1}(\mathbf{z}_i) \tag{9}$$

where $Q \in \mathbb{R}^{h\times w\times c}$ and $Conv_{1\times1}$ represents a 1×1 convolution operator to compensate for the depth difference with the intermediate feature layer. Then, different linear projections of the input feature map $X_i \in \mathbb{R}^{h\times w\times c}$ are constructed to obtain K and V in the cross-attention mechanism:

$$\begin{aligned} K &= Conv_{1\times1}(X_i) \\ V &= Conv_{1\times1}(X_i) \end{aligned} \tag{10}$$

where $K \in \mathbb{R}^{h\times w\times c}$ and $V \in \mathbb{R}^{h\times w\times c}$. The output feature map X_{iout} can be obtained:

$$X_{iout} = Softmax(\frac{Q\cdot K^T}{\sqrt{d_k}})\cdot V \tag{11}$$

where d_k is the number of columns of matrix Q.

Moreover, in real-world scenarios where non-uniform blur is common, single-scale features may not adequately handle complex blurry situations [38]. Unlike the approach in [38], which uses downsampling for multi-scale extraction, our wavelet transform inherently offers multi-scale advantages without information loss. Consequently, our proposed denoising process integrates multi-scale features $\mathbf{z}_i$ $(i = 1, 2, 3)$ with intermediate features X_i in a hierarchical manner to achieve better outcomes.

3.3 Network Training

In our approach, we jointly optimize PFRNet and DMDNet. For optimizing PFRNet, we introduce a loss function $\mathcal{L}_{DWT}$ in wavelet domain, which is the mean square error between the magnitudes of the DWT coefficients of the ground-truth image and predicted image defined by:

$$\begin{aligned} \mathcal{L}_{DWT} = \sum_{i=1}^{L} \mathbb{E}[&\left\|\hat{A}_i - A_i\right\|^2 + \left\|\hat{V}_i - V_i\right\|^2 \\ &+ \left\|\hat{H}_i - H_i\right\|^2 + \left\|\hat{D}_i - D_i\right\|^2] \end{aligned} \tag{12}$$

where A_i, V_i, H_i, D_i are the sub-bands of the ground-truth image at the i-th downsampling level, and

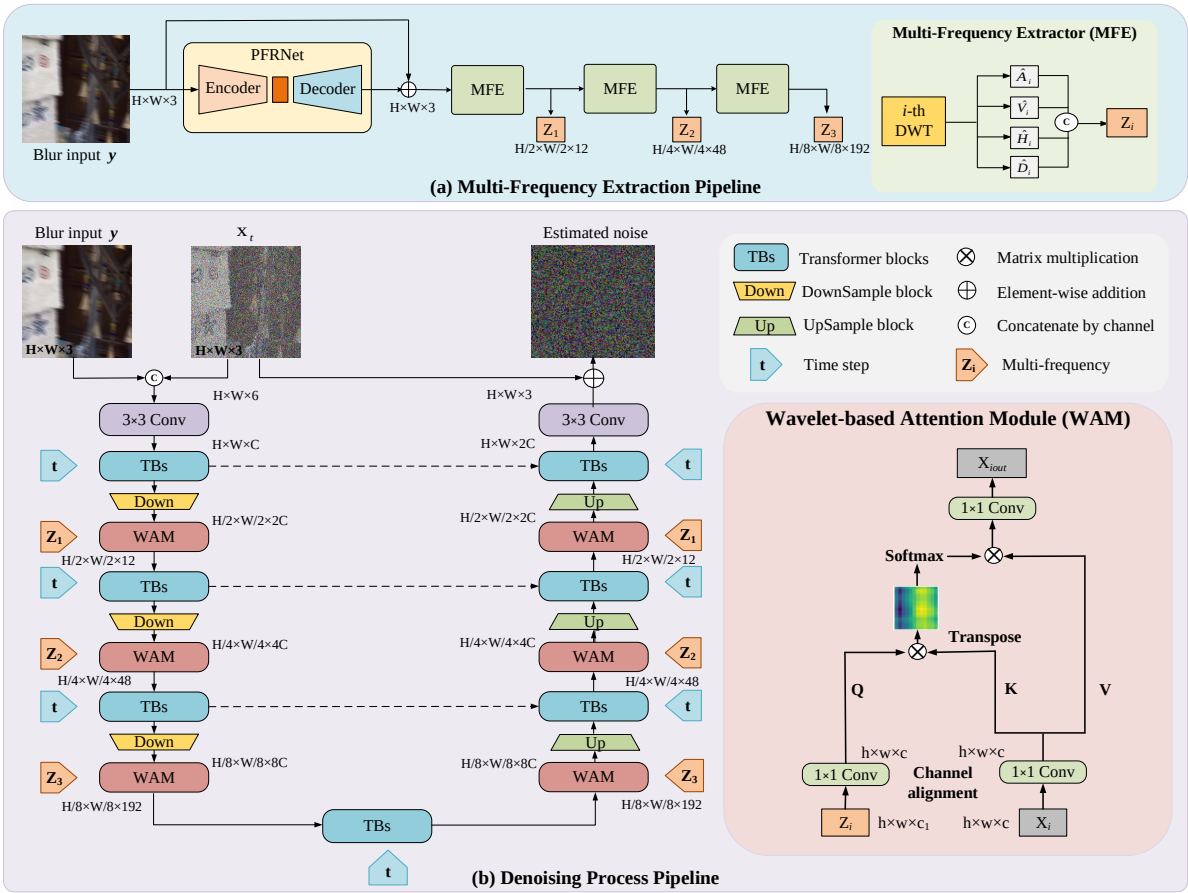


Fig. 5: The overall framework of the denoising process. (a) Multi-Frequency Extractor extracts the coefficients $\{\hat{A}_i, \hat{V}_i, \hat{H}_i, \hat{D}_i\}$ of the PFRNet prediction at the i-th level using the 2D-DWT. (b) Wavelet-based Attention Module is hierarchically integrated into the Transformer-based denoising network.

$\hat{A}_i, \hat{V}_i, \hat{H}_i, \hat{D}_i$ are the sub-bands of the predicted image at the i-th downsampling level. L denotes the number of downsampling levels, which is set to 3 in this work.

For optimizing DMDNet, we employ a two-stage approach involving noise loss and content loss. The noise loss $\mathcal{L}_{noise}$ serves as a constraint on noise:

$$\mathcal{L}_{noise} = \mathbb{E}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\cdot\epsilon, \mathbf{y}, \mathbf{z}, t\right)\right\|^2\right] \tag{13}$$

The content loss $\mathcal{L}_{content}$ acts as a constraint on the sampled restoration results and their corresponding ground truths, defined as

$$\mathcal{L}_{content} = \lambda\left(1 - \text{SSIM}\left(\mathbf{x}_0, \mathbf{x}\right)\right) + (1-\lambda)\|\mathbf{x}_0 - \mathbf{x}\|_1 \tag{14}$$

where $\mathbf{x}_0$ denotes the generated sampled restoration results; λ is empirically set to 0.84, as in [24].

Finally, we employ a coarse-to-fine joint training strategy. During coarse training, the total loss function is defined as $\mathcal{L} = \mathcal{L}_{DWT} + \mathcal{L}_{noise}$, focusing on enhancing noise constraints within the DM. In fine training, the total loss becomes $\mathcal{L} = \mathcal{L}_{DWT} + \mathcal{L}_{content}$, emphasizing content fidelity between sampled restorations and ground truths. Fine training aims to enhance the recovery quality of DMs, addressing inaccuracies in noise estimation from the coarse training stage.

4 Experiments

4.1 Experimental Settings

Data and Performance metrics. Following previous image deblurring methods, we evaluate our approach on commonly public image deblurring datasets, including the GoPro dataset [31], the HIDE dataset

Table 1: Evaluation results on the GoPro [31] and HIDE [44] datasets with GoPro [31] trained models. The CNN, TF, and DM categories denote methods primarily based on convolutional neural networks, Transformers, and diffusion models, respectively. These categories are determined by the core technical innovation of each method. The best and second best are marked in **bold** and underlined, respectively. The up-arrow ↑ indicates the larger value achieved, the better performance is, while the down-arrow ↓ indicates the smaller, the better.

Methods		GoPro [31]			HIDE [44]		
		PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
CNN	DeblurGAN-v2 [13]	29.08	0.876	0.119	27.51	0.852	0.160
	SRN [32]	30.25	0.901	0.130	28.36	0.875	0.154
	DBGAN [45]	31.18	0.916	0.112	28.94	0.889	0.145
	MIMO [6]	31.72	0.924	0.104	29.28	0.896	0.141
	MPRNet [46]	32.66	0.936	0.091	30.96	0.918	0.116
	DGUNet [47]	32.70	0.937	0.094	30.96	0.919	0.125
	HINet [48]	32.77	0.936	0.090	30.32	0.910	0.122
	LaKDNet [49]	33.35	0.943	0.084	31.21	0.923	0.110
	NAFNet [34]	**33.71**	**0.947**	0.081	31.32	0.924	0.109
TF	Restormer [11]	32.92	0.939	0.086	31.22	0.922	0.110
	Uformer [36]	33.05	0.941	0.089	30.89	0.920	0.116
	Stripformer [9]	33.08	0.941	0.088	31.03	0.918	0.108
DM	HI-Diff [38]	33.33	0.945	0.082	**31.46**	**0.925**	0.112
	IR-SDE [50]	30.63	0.901	0.069	25.76	0.825	0.130
	C2F-DFT [24]	31.96	0.928	0.100	30.52	0.915	0.118
	Ours	32.52	0.935	**0.066**	30.89	0.920	**0.106**

Table 2: Evaluation results on the RealBlur-R [51] and RealBlur-J [51] with GoPro [31] trained models. The best and second best are marked in **bold** and underlined, respectively.

Methods		RealBlur-R [51]			RealBlur-J [51]		
		PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
CNN	DeblurGAN-v2 [13]	35.10	0.934	0.086	28.69	0.865	0.144
	SRN [32]	35.41	0.929	0.087	28.49	0.863	0.160
	DBGAN [45]	33.78	0.908	0.135	24.93	0.745	0.295
	MIMO [6]	35.47	0.946	0.089	27.75	0.836	0.191
	MPRNet [46]	35.98	0.952	0.075	28.69	0.872	0.158
	DGUNet [47]	35.95	0.952	0.071	28.75	0.875	0.152
	HINet [48]	35.75	0.949	0.078	28.17	0.849	0.173
	LaKDNet [49]	35.90	0.954	0.061	28.77	0.878	0.152
	NAFNet [34]	35.97	0.951	0.068	28.31	0.856	0.167
TF	Restormer [11]	36.19	0.957	0.061	28.96	0.878	0.155
	Uformer [36]	36.18	0.956	0.062	29.08	0.885	0.146
	Stripformer [9]	36.07	0.954	0.063	28.82	0.876	0.145
DM	HI-Diff [38]	36.28	**0.958**	0.060	29.15	0.889	0.146
	IR-SDE [50]	33.96	0.918	0.114	24.21	0.729	0.267
	C2F-DFT [24]	36.34	0.957	0.064	28.89	0.876	0.154
	Ours	**36.37**	**0.958**	**0.058**	**29.23**	**0.891**	**0.141**

[44], the RealBlur-R dataset [51], and the RealBlur-J dataset [51]. The GoPro dataset [31] contains 3214 image pairs of size 1280 × 720, with 2103 for training and 1111 for testing. The HIDE dataset [44] comprises 2025 image pairs of the same size for testing. The RealBlur-R dataset [51] and RealBlur-J dataset [51] are two real-world datasets mainly consisting of motion blurred low light scenes, and each dataset includes 980 test images. Our method is trained on the GoPro dataset [31], and subsequently evaluated on the HIDE dataset [44], the RealBlur-R dataset [51] and RealBlur-J dataset [51] to gauge its generalization capacity across different distributions. Two objective metrics (Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [52]) and a perceptual metric (Learned Perceptual Image Patch Similarity (LPIPS)) [53] are employed to evaluate the performance of the proposed method.

Implementation Details. Our implementation is carried out using PyTorch and running on a server equipped with an NVIDIA Tesla V100 GPU for both training and testing of the proposed image deblurring

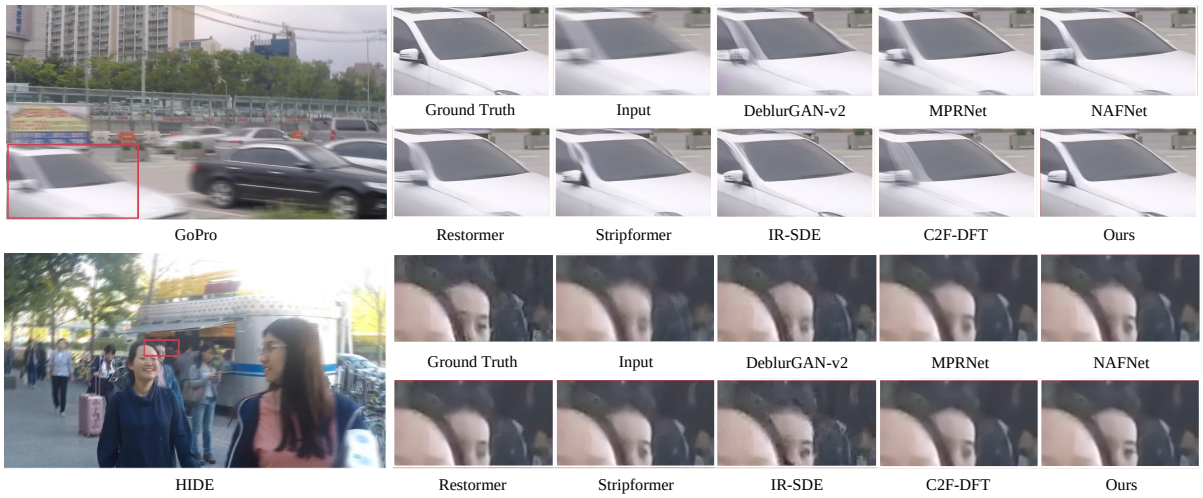


Fig. 6: Qualitative comparison of our method and competitive methods on GoPro [31] and HIDE [44] test sets.

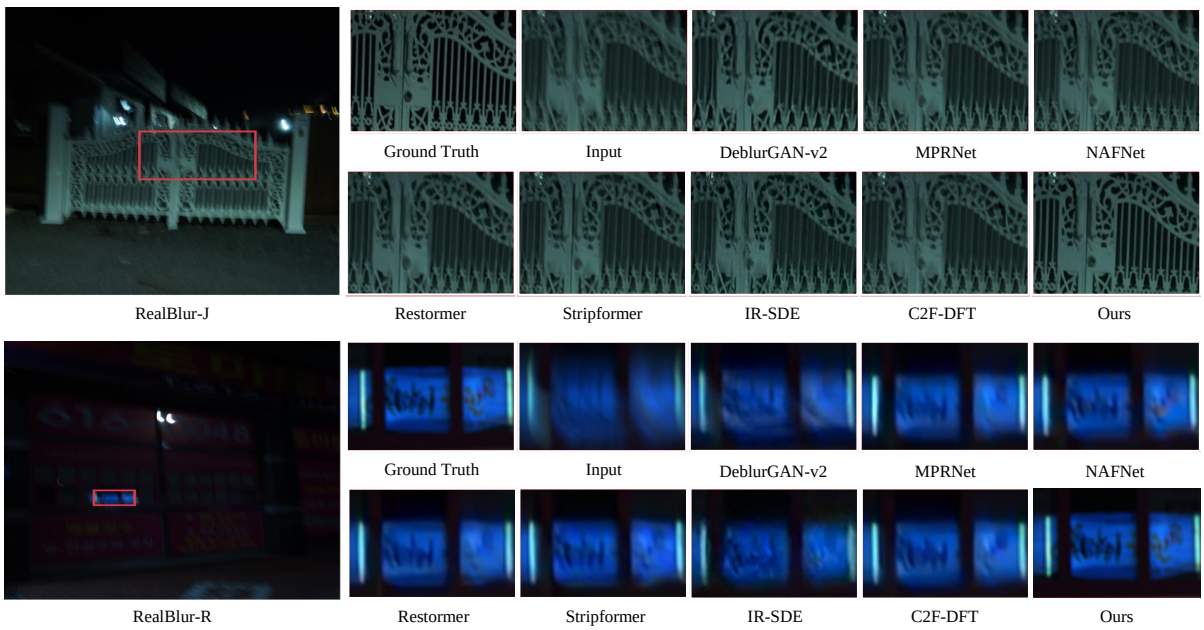


Fig. 7: Qualitative comparison of our method and competitive methods on RealBlur-R [51] and RealBlur-J [51] test sets.

model. DMDNet consists of a four-level U-shaped structure with transformer blocks, where the number of transformer blocks is [4,6,6,8], attention heads are [1,2,4,8], and the number of channels is [48, 96, 192, 384]. In the diffusion process, we set the noise variances to linearly increase from $\beta_1 = 1e^{-4}$ to $\beta_2 = 2e^{-2}$. The maximum time step T is set to 1000, with a sampling step S of 3. We warm start the training with only wavelet loss, and linearly increase the weight of the noise loss to 1. During the coarse training phase, we employ the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay $1e^{-4}$. The learning rate is set to $3e^{-4}$ and gradually reduced to $1e^{-5}$ using cosine annealing [54]. Additionally, we incorporate a progressive learning strategy to capture more contextual information. Specifically, we progressively update the patch size and batch size pairs to $[(64^2, 64), (128^2, 16), (256^2, 4)]$ every 10K iterations during the coarse training phase, totaling 1600K iterations. In the subsequent fine training phase, the

learning rate is set to $1e^{-5}$ and gradually reduced to $1e^{-7}$ using cosine annealing [54]. Additionally, we progressively update the patch size and batch size pairs to [(64^2, 16), (128^2, 4), (256^2, 1)] every 10K iterations, with a total of 500K fine training iterations.

4.2 Experimental Results

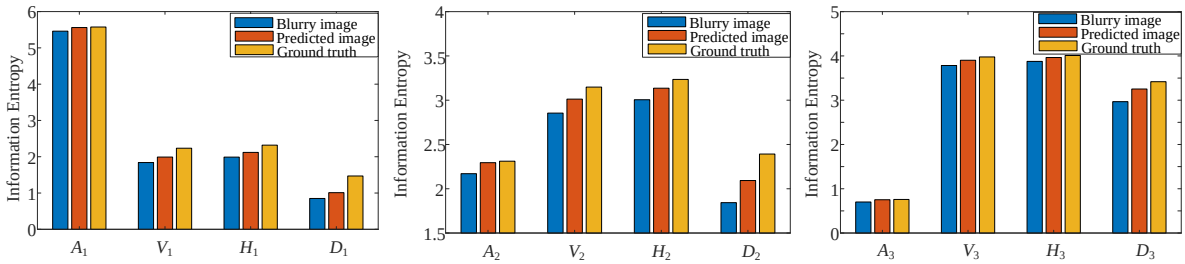
In this section, we compare our proposed method with existing state-of-the-art methods. The methods are categorized based on their core technical innovation: (1) **CNN-based methods**: DeblurGAN-v2 [13], SRN [32], DBGAN [45], MIMO [6], MPRNet [46], DGUNet [47], HINet [48], LaKDNet [49], NAFNet [34]; (2) **Transformer-based methods**: Restormer [11], Uformer [36], Stripformer [9]; (3) **Diffusion-based methods**: IR-SDE [50], C2F-DFT [24], HI-Diff [38]. Although HI-Diff [38] combines diffusion models with Transformers, we classify it into diffusion-based methods because its key innovation lies in applying the concept of the diffusion model. Specifically, HI-Diff [38] leverages the power of diffusion models to generate informative priors, which are then integrated into a Transformer architecture to enhance the details of the deblurred image. We test the above methods using the code and weights provided by their respective authors. For fair comparisons, all these methods are trained on the GoPro dataset [31] and are subsequently applied directly to the HIDE [44], RealBlur-R [51] and RealBlur-J [51].

Quantitative Results. The quantitative results on the GoPro [31] and HIDE [44] datasets are presented in Tab. 1. While our method slightly lags in distortion metrics (PSNR/SSIM), it excels in the perception metric LPIPS. Notably, our method significantly outperforms diffusion model-based methods such as IR-SDE [50] and C2F-DFT [24]. HI-Diff [38] integrates diffusion priors into transformer-based architectures, leveraging the distortion accuracy of regression-based methods. Consequently, HI-Diff achieves distortion measurements similar to transformer methods, outperforming our method in this aspect. Regarding the perceptual metric LPIPS, where CNN-based, Transformer-based, and previous diffusion model-based methods underperform, our method achieves the lowest scores on both the GoPro dataset [31] and HIDE dataset [44], underscoring our model's ability to generate visually pleasing restored images. On both datasets, our method improves the perceptual-based metric LPIPS by 0.003 and 0.002, respectively, compared to the runner-up method.

We further evaluate the GoPro-trained model on the real-world dataset: RealBlur-R [51] and RealBlur-J [51]. As reported in Tab. 2, our method achieves state-of-the-art performance, demonstrating robust generalization capabilities on these real-world datasets. Compared to NAFNet [34], which is the best-performing method on the GoPro dataset [31] among the existing methods, our approach demonstrates improvements of 0.4 dB on the RealBlur-R dataset [51] and 0.92 dB on the RealBlur-J dataset [51]. On the RealBlur-R dataset [51], our method improves PSNR by 0.03 db compared to the second-best method C2F-DFT [24]. Similarly, on the RealBlur-J dataset [51], our method surpasses the second-best method HI-Diff [38] by 0.08dB in PSNR and 0.002 in SSIM. Furthermore, our method gets the lowest average scores in the perceptual metric LPIPS, indicating its superior generalization ability to unseen real-world scenes.

Based on the results in Tabs. 1 and 2, our model exhibits slightly lower PSNR and SSIM scores on the GoPro [31] test set and the HIDE [44] dataset compared to some baselines. This is due to the nature of diffusion models, which prioritize generating realistic images with rich visual details over pixel-wise fidelity. Despite these lower PSNR and SSIM scores, our model achieves the best LPIPS scores across all datasets, indicating a trade-off between enhanced perceptual quality and pixel-wise accuracy. The GoPro [31] and HIDE [44] datasets are synthetic, generally noise-free, and have limited variation in motion blur. In contrast, the RealBlur-R and RealBlur-J datasets [51] contain complex real-world noise and blur. Notably, our model significantly outperforms other methods on the more challenging RealBlur-R and RealBlur-J datasets [51]. This superior performance can be attributed to our model's iterative refinement process, which allows for detailed and gradual correction of complex noise and blur. Additionally, our model leverages hierarchical wavelet-derived multi-frequency information as conditions to guide the denoising process, improving its ability to restore intricate details in challenging real-world conditions.

Qualitative Results. We present qualitative comparisons between our method and state-of-the-art methods (DeblurGAN-v2 [13], MPRNet [46], NAFNet [34], Restormer [11], Stripformer [9], IR-SDE [50], C2F-DFT [24]) on the GoPro [31], HIDE [44], RealBlur-R [51], and RealBlur-J [51] datasets, as illustrated in Fig. 6 and Fig. 7. It can been seen



(a) Wavelet coefficients (b) Wavelet coefficients (c) Wavelet coefficients

Fig. 8: Information entropy of wavelet coefficients at three scales for three types of inputs (blurry image, PFRNet-predicted image, Ground truth) on the GoPro dataset [31].

Table 3: Ablation studies of the PFRNet on GoPro dataset [31]. The options include "N/A" for no CNN used, '-S', '-M', and '-L' for small, middle, and large networks. The best results are highlighted in **bold**.

	Network Size	Params (M)	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
N/A	-	-	30.92	0.911	0.117
CNN-S	ch=16	3.69	32.52	0.935	**0.066**
CNN-M	ch=32	14.53	32.54	0.935	0.067
CNN-L	ch=64	57.67	**32.55**	**0.936**	0.069

that the proposed method recovers more precise textures and sharper edges than the existing methods. In the GoPro sample, our propsed method restores the rearview mirror of the car with more sharper edges devoid of artifacts, compared with other existing models. In the HIDE sample, competing methods struggle to reconstruct human eyebrows and eyes accurately, while our method successfully recovers sharp textures. Remarkably, our method even surpasses Ground Truth in visual quality. In the RealBlur-J sample, the most of existing state-of-the-art methods introduce adverse effects in real-world images, resulting in even blurrier outcomes than the original blurry ones. In contrast, our method produces sharper results with finer structures. In the RealBlur-R sample, our method effectively restores the billboard number with reduced blur, while other algorithms struggle to recover the numbers on the billboard. These qualitative results show that our method is a robust deblurring diffusion model, particularly in handling real scenes.

4.3 Ablation Study

In this section, we conduct a series of ablation studies to analyze the impacts of various components and training strategy configurations.

Effects of PFRNet. In Tab. 3, we evaluate the performance with and without PFRNet, as well as with different network sizes. Directly using blurry images without PFRNet for initial estimation leads to notably inferior performance compared to employing PFRNet. This disparity primarily arises from the limited frequency information available in blurry images. Fig. 8 illustrates the average information entropy of wavelet coefficients at three scales for different image types (blurry, PFRNet-predicted, Ground truth) on the GoPro dataset [31], confirming that PFRNet-predicted images retain richer frequency details. We experiment with different channel numbers (ch = 16, 32, 64), denoting networks as S, M, L. Due to the frequency constraints during training, varying network sizes had minimal impact on our model's performance. Even with larger networks, the SSIM performance showed marginal improvement (0.001), indicating that the smallest network suffices for our requirements.

Table 4: Ablation studies of the wavelet coefficients extracted by MFE on GoPro dataset [31] and RealBlur-J dataset [51]. The best results are highlighted in **bold**.

	GoPro [31]			RealBlur-J [51]		
	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
LF ($[A]$)	31.54	0.920	0.110	27.76	0.835	0.195
HF ($[V, H, D]$)	31.67	0.922	0.108	28.15	0.843	0.172
MF ($[A, V, H, D]$)	**32.52**	**0.935**	**0.066**	**29.23**	**0.891**	**0.141**

Table 5: Ablation studies of the effectiveness of the proposed WAM on GoPro dataset [31] and RealBlur-J dataset [51]. The best results are highlighted in **bold**.

	GoPro [31]			RealBlur-J [51]		
	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Basline	31.72	0.926	0.101	28.76	0.875	0.152
Basline + WAM	**32.52**	**0.935**	**0.066**	**29.23**	**0.891**	**0.141**

Effects of MFE. To further explore what wavelet coefficients extracted by MFE should be used in the

Fig. 9: Visual results without and with WAM. Our method with WAM generates sharper textures and more complete structures compared to the Baseline.

Fig. 10: Coarse and fine training visual results. Coarse training fails to yield promising results due to inaccurate noise estimation, while fine training effectively enhances restoration quality.

Table 6: Ablation studies of the coarse-to-fine joint training strategy on GoPro [31] test set. The **above** part illustrates the joint training of PFRNet and DMD-Net, comparing the results with those from split training. The **below** part presents coarse-to-fine training, showing the results of different training iterations. The best results are highlighted in **bold**.

Training Strategy	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Split training	31.15	0.915	0.112
Joint training	**32.52**	**0.935**	**0.066**
Coarse 1600k	31.45	0.925	0.108
Coarse 2100k	31.51	0.926	0.105
Coarse 1600k + Fine 500k	**32.52**	**0.935**	**0.066**

diffusion model, several choices for them with corresponding quantitative results are shown in Tab. 4. According to them, we can observe that the restoration performance reaches the best when extracting the multi-frequency (MF) coefficients that contain both low-frequency (LF) and high-frequency (HF) components. This optimal performance can be attributed to the integration of global structural and local detail information provided by these coefficients, which results in a richer feature representation and minimizes information loss.

Effects of WAM. We further validate the effectiveness of the proposed WAM. We construct a baseline model, denoted as Baseline, which is a standard DM condition on blurred images. Upon this Baseline, we introduce the WAM (Baseline + WAM). It can be seen from Tab. 5 that utilizing the WAM yields improvements of 0.8 dB, 0.009 and 0.035 in terms of PSNR, SSIM and LPIPS on the GoPro dataset [31]. In addition, on the RealBlur-J dataset [51], our method with WAM exhibit a significant performance improvement of 0.47 dB, 0 .015, and 0.011 in PSNR, SSIM, and LPIPS, respectively. Furthermore, we provide a

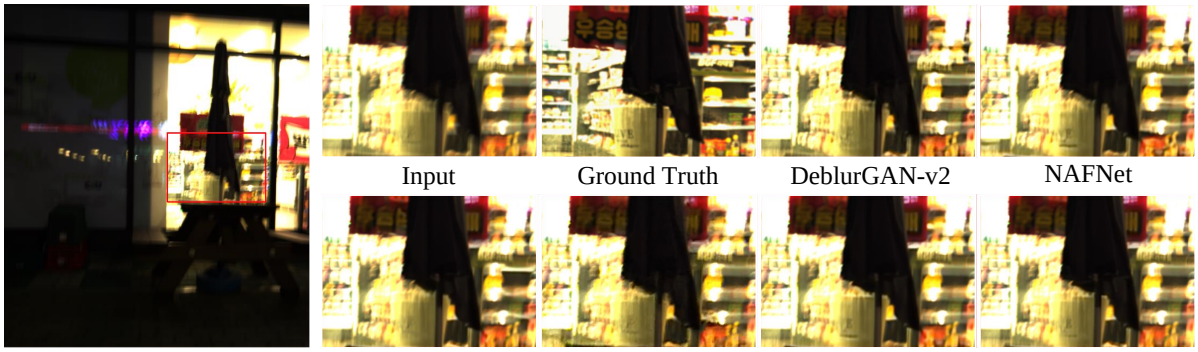


Fig. 11: Failure case in low-light conditions from Realblur-R [51] across multiple deblurring methods. including DeblurGAN-v2 [13], NAFNet [34], Restormer [11], IR-SDE [50], C2F-DFT [24], and Ours.

visual comparison in Fig. 9. It can be observed that our method with WAM generates more complete wheel structures and finer details of flowers compared to the Baseline. The main reason is that the proposed WAM can adaptively focus on frequency information, guiding the diffusion model to recover more low-frequency structures and fine-grained high-frequency details.

Effects of Training Strategy. We employ a coarse-to-fine joint training strategy to train our model, which raises questions about their effectiveness on restoration quality. For comparison, we also conduct a split training approach, where the DMD-Net is optimized independently while the PFRNet is pre-trained. For the PFRNet training, we use the same experimental setup as described in Sec. 4.1, but set the total number of iterations to 100K. As shown in Tab. 6, joint training significantly outperforms split training by 1.37 dB in PSNR, highlighting the importance of the joint training strategy. Furthermore, we investigate the impact of training from coarse to fine. It is evident from Tab. 6 that additional iterations during coarse training yield only a minor improvement of 0.06 dB in PSNR. In contrast, incorporating fine training after completing the coarse training results in a substantial PSNR improvement of 1.07 dB. Fig. 10 show an example of the comparison between coarse and fine training. We find that coarse training fails to restore promising results, while our fine training significantly improves the visual restoration quality. This shows that the fine training, which constrains the sampled restored results with ground truth instead of noise to optimize the diffusion model, can help the model learn more complex degraded information, facilitating better restoration.

4.4 Limitations

Our proposed method is effective in image deblurring, but it has also challenges in extremely low-light environments where significant information loss complicates restoration. Furthermore, the exclusive training of our method on the GoPro dataset [31] restricts its adaptability to a variety of real-world situations. We recognize that the quality and realism of the training dataset directly influence our model's deblurring capabilities. As depicted in Fig. 11, the deblurring methods fails to restore the images under low-light conditions in the Realblur-R dataset [51]. Consequently, we advocate for the availability of large-scale and diverse training data to bolster our method's performance. Additionally, our model's capacity is inherently constrained to generalize solely to the deblurring task observed during training. Investigating the applicability of our proposed approach to other image restoration tasks remains a promising avenue for future research.

5 Conclusion

In this work, we propose a hierarchical wavelet-guided diffusion model for image deblurring. Specifically, we introduce the PFRNet to generate a coarse estimate and employ multi-level wavelet transforms to extract multi-frequency information as guidance. Additionally, we present a wavelet-based attention module to integrate this guidance with intermediate features of the DM across multiple layers. Our model effectively captures frequency information at multiple scales to

improve the quality of restored image in complex blurring scenarios. Furthermore, we proposed a coarse-to-fine joint optimization strategy for both the PFRNet and DMDNet to ensure optimal performance. Comprehensive experiments on real-world blur datasets show that our proposed method achieves superior performance compared with existing state-of-the-art methods.

CRediT authorship contribution statement

Xiaopan Li: Conceptualization, Methodology, Validation, Software, Writing - original draft. **Shiqian Wu:** Formal analysis, Validation, Writing - review & editing. **Xin Yuan:** Writing - review & editing. **Shoulie Xie:** Conceptualization, Methodology, Writing - review & editing. **Sos Agaian:** Supervision, writing - review & editing.

Declaration of Competing Interest

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

Data Availability

The available online datasets in this paper are https://github.com/subeeshvasu/Awesome-Deblurring.

References

[1] Sheng, B., Li, P., Fang, X., Tan, P., Wu, E.: Depth-aware motion deblurring using loopy belief propagation. IEEE Transactions on Circuits and Systems for Video Technology **30**(4), 955–969 (2020)

[2] Wen, Y., Chen, J., Sheng, B., Chen, Z., Li, P., Tan, P., Lee, T.-Y.: Structure-aware motion deblurring using multi-adversarial optimized cyclegan. IEEE Transactions on Image Processing **30**, 6142–6155 (2021)

[3] Chen, L., Fang, F., Wang, T., Zhang, G.: Blind image deblurring with local maximum gradient prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1742–1750 (2019)

[4] Li, X., Wu, S., Xie, S., Agaian, S.: Dynamic-clustering extreme intensity prior based blind image deblurring. Journal of Mathematical Imaging and Vision **66**(1), 22–36 (2024)

[5] Feng, X., Tan, J., Ge, X., Liu, J., Hu, D.: Blind image deblurring via weighted dark channel prior. Circuits, Systems, and Signal Processing **42**(9), 5478–5499 (2023)

[6] Cho, S.-J., Ji, S.-W., Hong, J.-P., Jung, S.-W., Ko, S.-J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4641–4650 (2021)

[7] Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5978–5986 (2019)

[8] Feng, Z., Zhang, J., Ran, X., Li, D., Zhang, C.: Ghost-unet: multi-stage network for image deblurring via lightweight subnet learning. The Visual Computer, 1–15 (2024)

[9] Tsai, F.-J., Peng, Y.-T., Lin, Y.-Y., Tsai, C.-C., Lin, C.-W.: Stripformer: Strip transformer for fast image deblurring. In: Proceedings of the European Conference on Computer Vision, pp. 146–162 (2022)

[10] Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5886–5895 (2023)

[11] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5728–5739 (2022)

[12] Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8183–8192 (2018)

[13] Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8878–8887 (2019)

[14] Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics **42**(6), 1–14 (2023)

[15] Croitoru, F.-A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10850–10869 (2023)

[16] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 6840–6851 (2020)

[17] Lu, X., Liu, F., Rong, Y., Chen, Y., Xiong, S.: Makeupdiffuse: a double image-controlled diffusion model for exquisite makeup transfer. The Visual Computer, 1–17 (2024)

[18] Jin, P., Li, H., Cheng, Z., Li, K., Ji, X., Liu, C., Yuan, L., Chen, J.: Diffusionret: Generative text-video retrieval with diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2470–2481 (2023)

[19] Xu, X., Yuan, X., Wang, Z., Zhang, K., Hu, R.: Rank-in-rank loss for person re-identification. ACM Transactions on Multimedia Computing, Communications, and Applications **18**(2s), 1–21 (2022)

[20] Shang, S., Shan, Z., Liu, G., Zhang, J.: Resdiff: Combining CNN and diffusion model for image super-resolution. arXiv preprint arXiv:2303.08714 (2023)

[21] Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A.G., Milanfar, P.: Deblurring via stochastic refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 16293–16303 (2022)

[22] Ren, M., Delbracio, M., Talebi, H., Gerig, G., Milanfar, P.: Multiscale structure guided diffusion for image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10721–10733 (2023)

[23] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)

[24] Wang, L., Yang, Q., Wang, C., Wang, W., Pan, J., Su, Z.: Learning a coarse-to-fine diffusion transformer for image restoration. arXiv preprint arXiv:2308.08730 (2023)

[25] Niu, A., Zhang, K., Pham, T.X., Sun, J., Zhu, Y., Kweon, I.S., Zhang, Y.: Cdpmsr: Conditional diffusion probabilistic models for single image super-resolution. In: 2023 IEEE International Conference on Image Processing, pp. 615–619 (2023). IEEE

[26] Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-cnn for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 773–782 (2018)

[27] Huang, Y., Huang, J., Liu, J., Yan, M., Dong, Y., Lyu, J., Chen, C., Chen, S.: Wavedm: Wavelet-based diffusion models for image restoration. IEEE Transactions on Multimedia (2024)

[28] Li, J., Cheng, B., Chen, Y., Gao, G., Shi, J., Zeng, T.: Ewt: Efficient wavelet-transformer for single image denoising. Neural Networks **177**, 106378 (2024)

[29] Mao, X., Liu, Y., Liu, F., Li, Q., Shen, W., Wang, Y.: Intriguing findings of frequency selection for image deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1905–1913 (2023)

[30] Liu, Y., Fang, F., Wang, T., Li, J., Sheng, Y., Zhang, G.: Multi-scale grid network for image deblurring with high-frequency guidance. IEEE Transactions on Multimedia **24**, 2890–2901 (2021)

[31] Nah, S., Hyun Kim, T., Mu Lee, K.: 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)

[32] Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8174–8182 (2018)

[33] Ji, S.-W., Lee, J., Kim, S.-W., Hong, J.-P., Baek, S.-J., Jung, S.-W., Ko, S.-J.: Xydeblur: divide and conquer for single image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 17421–17430 (2022)

[34] Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Proceedings of the European Conference on Computer Vision, pp. 17–33 (2022)

[35] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[36] Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 17683–17693 (2022)

[37] Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. arXiv preprint arXiv:2303.09472 (2023)

[38] Chen, Z., Zhang, Y., Liu, D., Gu, J., Kong, L., Yuan, X., *et al.*: Hierarchical integration diffusion model for realistic image deblurring. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1–12 (2024)

[39] Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 10346–10357 (2023)

[40] Min, C., Wen, G., Li, B., Fan, F.: Blind deblurring via a novel recursive deep cnn improved by wavelet transform. IEEE Access **6**, 69242–69252 (2018)

[41] Zou, W., Jiang, M., Zhang, Y., Chen, L., Lu, Z., Wu, Y.: Sdwnet: A straight dilated network with wavelet transformation for image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1895–1904 (2021)

[42] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[43] Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687 (2022)

[44] Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5572–5581 (2019)

[45] Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., Li, H.: Deblurring by realistic blurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2737–2746 (2020)

[46] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 14821–14831 (2021)

[47] Mou, C., Wang, Q., Zhang, J.: Deep generalized unfolding networks for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 17399–17410 (2022)

[48] Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 182–192 (2021)

[49] Ruan, L., Bemana, M., Seidel, H.-p., Myszkowski, K., Chen, B.: Revisiting image deblurring with an efficient convnet. arXiv preprint arXiv:2302.02234 (2023)

[50] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. arXiv preprint arXiv:2301.11699 (2023)

[51] Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: Proceedings of the European Conference on Computer Vision, pp. 184–201 (2020). Springer

[52] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)

[53] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)

[54] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: Proceedings of the International Conference on Learning Representations, pp. 1–16 (2017)