Learning Diffusion Texture Priors for Image Restoration

Tian Ye[1] Sixiang Chen[1] Wenhao Chai[2]
Zhaohu Xing[1] Jing Qin[3] Ge Lin[1,4] Lei Zhu[1,4*]

[1] The Hong Kong University of Science and Technology (Guangzhou) [2] University of Washington
[3] The Hong Kong Polytechnic University [4] The Hong Kong University of Science and Technology

Abstract

Diffusion Models have shown remarkable performance in image generation tasks, which are capable of generating diverse and realistic image content. When adopting diffusion models for image restoration, the crucial challenge lies in how to preserve high-level image fidelity in the randomness diffusion process and generate accurate background structures and realistic texture details. In this paper, we propose a general framework and develop a Diffusion Texture Prior Model (DTPM) for image restoration tasks. DTPM explicitly models high-quality texture details through the diffusion process, rather than global contextual content. In phase one of the training stage, we pre-train DTPM on approximately 55K high-quality image samples, after which we freeze most of its parameters. In phase two, we insert conditional guidance adapters into DTPM and equip it with an initial predictor, thereby facilitating its rapid adaptation to downstream image restoration tasks. Our DTPM could mitigate the randomness of traditional diffusion models by utilizing encapsulated rich and diverse texture knowledge and background structural information provided by the initial predictor during the sampling process.

1. Introduction

Image Restoration tasks [37, 43, 70, 71] usually are ill-posed problems whose solutions are not unique. While existing methods [8, 70, 71] have achieved notable breakthroughs, they typically employ direct regression models to produce deterministic results. A persistent challenge is that these deterministic models [4, 70, 71] frequently yield unsatisfactory fine-grained details (See Figure 2), as they are trained to minimize pixel-level error, aligning output with ground truth by Norm-based losses [30]. The emergence of Diffusion Models (DMs) [14, 16, 51, 55] has revolutionized image generation, yielding realistic images replete with fine details. This advancement motivates our investigation into diffusion-based methods for image restoration, aiming to leverage conditional DMs [17, 45, 51] to produce perceptually appealing results that closely mimic natural, clean images.

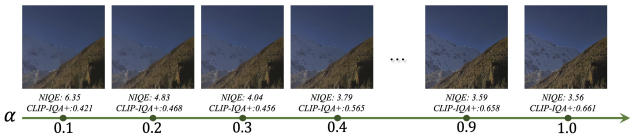

Figure 1. We use a coefficient α ranging from 0.1 to 1 to control the intensity of texture removal. As α increases, leading to richer textures from left to right in the images, we can observe a corresponding improvement in the scores of no-reference image quality metrics, specifically NIQE [42] and CLIP-IQA+ [57]. Regarding the NIQE, lower scores indicate better quality, while for CLIP-IQA+, higher scores represent better visual quality. **This indicates that the quality of the image's texture details have a significant impact on the visual perceptual quality of the image.**

However, directly applying diffusion image generation techniques [44, 51, 55] to image restoration tasks is often impractical [36, 45]. The high content fidelity required in image restoration clashes with the stochastic nature of diffusion models [44, 55]. A plausible solution involves integrating physical degradation models [5, 69] with neural networks or crafting hand-designed priors [20, 47] to mitigate the inherent randomness of the diffusion paradigm. While several previous studies [17, 68, 69] have adopted this approach with success, they lack versatility and show limited generalization capability on real-world scenes, as physical degradation models do not fully encompass real-world degradation scenarios [4, 66, 67]. Additionally, the iterative nature and complexity of the denoising process in diffusion models necessitate extensive data and lengthy training cycles for effective learning. The challenges and limitations outlined above necessitate a rethinking of the use of the diffusion paradigm in image restoration: *(i) To preserve high-level fidelity in restoration, we propose using a diffusion model to recover only texture layers. This*

*Lei Zhu (leizhu @ust.hk) is the corresponding author.

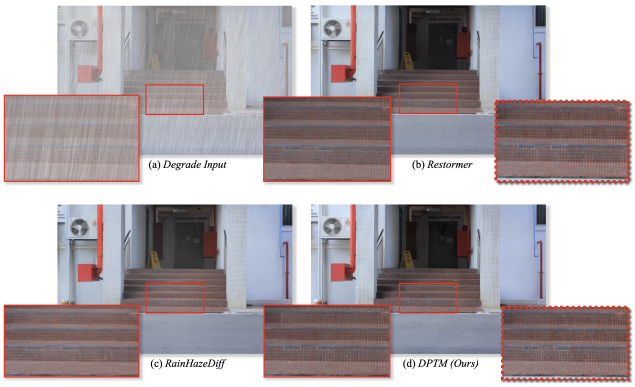

Figure 2. (a) The input degraded image. (b) The result of Restormer [70]. Restormer is a representative image restoration method that can only produce **deterministic results, which may lack promising details.** (c) The result of RainHazeDiff [45]. (d) The result of our method. The content within the dashed rectangle represents the corresponding ground-truth details. It's worth noting that our approach excels in generating finer details, thanks to the incorporation of *Diffusion Texture Priors* we proposed.

approach emphasizes and highlights the importance of fine textures in visual perception (See Figure 1). By focusing diffusion on texture recovery, we minimize randomness of diffusion process and leverage its strength in generating realistic details. (ii) Recognizing the diffusion paradigm as a potent representation learning tool that requires extensive data for convergence, we pre-trained our model, named the Diffusion Texture Priors Model (DTPM), on a large-scale dataset of high-quality images. This model, trained on approximately 55K samples, is tailored to reconstruct texture layers, embedding texture priors within the diffusion process. Our DTPM framework is characterized by: **(a)** A novel approach in diffusion modeling that focuses on texture layers through residual learning, enhancing efficiency in downstream restoration tasks. **(b)** The incorporation of Semantic Code as a constraint, guiding the diffusion model to produce semantically coherent textures. **(c)** To preserve the integrity of the diffusion texture priors, most parameters of the DTPM remain fixed after pre-training. We introduce a Conditional Degradation Adapter to facilitate rapid adaptation to various restoration tasks without the risk of catastrophic forgetting.

Our extensive experiments across **single-image defocus deblurring, image motion deblurring, desnowing, dehazing, deraining, and raindrop removal** demonstrate that our method not only surpasses strong regression models and recent diffusion-based models in *perceptual quality* but also exhibits *robust generalization* across diverse restoration tasks.

2. Related Works

Evolution in Image Restoration. Image restoration [4, 15, 19, 25, 39, 67, 71, 77], a pivotal research domain within the computer vision community, has witnessed significant evolution over recent years. Traditionally regarded as an ill-posed problem, image restoration poses unique challenges due to the multitude of potential solutions derivable from a single degraded image. Over the last decade, there has been a notable shift in the computer vision field from conventional, handcrafted prior-based approaches [20, 47] to a more data-centric, deep learning-driven paradigm [4, 15, 59, 70]. This transition has led to the development of advanced deep learning techniques, characterized by complicated network structures [65, 70, 71] and extensive training on carefully collected datasets. These approaches excel in learning mappings from degraded to clean data, effectively tackling the complexities of image restoration. Initially, deep learning methods [9–11, 66, 67] in this area tended to produce deterministic outcomes, which starkly contrasted with the fundamentally probabilistic nature of image restoration tasks [18, 46, 70]. This discrepancy has increasingly brought generative methods [17, 27, 46, 64, 76] to the forefront, highlighting their potential to address the challenges inherent in image restoration.

Generative image restoration methods. Recent developments [6, 7, 17, 41, 46, 61, 64, 74] have shed light on the fundamental flaws of earlier deep learning-driven image restoration techniques, specifically their (i) generate deterministic outputs and (ii) inability to restore high-quality, detailed textures, owing to their training focused on minimizing pixel-level discrepancies. Consequently, generative approaches are gaining traction. A notable strategy includes employing Generative Adversarial Networks (GANs) [24, 29, 48]. Nonetheless, GANs pose challenges due to their complex training procedures, instability during training, and occasional generation of unrealistic image attributes. The impressive efficacy of diffusion in image generation has introduced a promising avenue. Diffusion techniques adeptly tackle the aforementioned issues, offering a more stable training regimen and yielding images with realistic textures. Applications of diffusion-based methods in image super-resolution [52], shadow removal [17, 27], deblurring [60], and adverse weather removal [46] have shown promising results. However, there remains a need for comprehensive research to establish a consistent, reliable framework for diffusion-based image restoration tasks.

3. Our Framework

3.1. Preliminaries of Diffusion Models

Diffusion Models (DMs), as referenced in [22, 44], are a class of generative models that gradually infuse Gaussian noise into training data and then employ a denoiser to re-

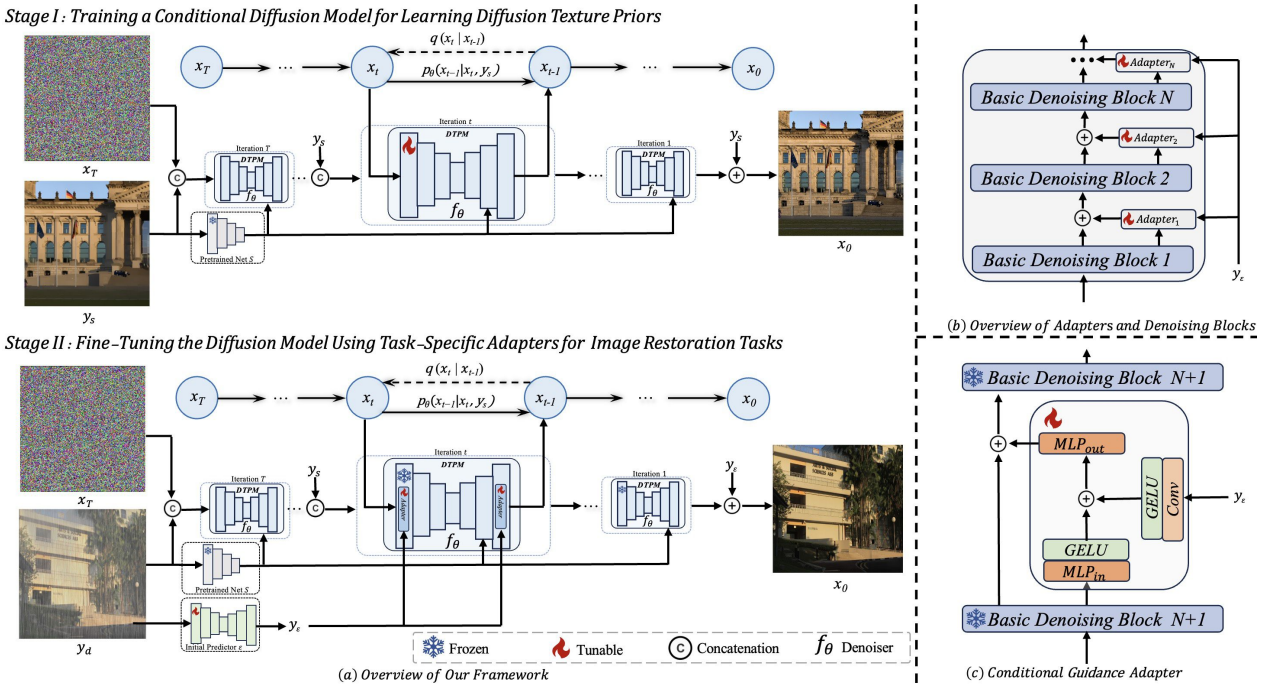

Figure 3. **The overview of Our DTPM Framework**. (a) Our method consists of two stages. *In the stage I*, under the guiding constraints of semantic code, the diffusion model learns texture layers through residual learning from a large amount of high-quality data, which allows us to encapsulate diverse and rich texture knowledge into the diffusion model. *In the stage II*, we fix most of the parameters of the trained diffusion model and insert Conditional Guidance Adapters between each layer for efficient fine-tuning and conditional guidance on image restoration tasks. (b) The overview of adapters and our denoising blocks.(c) The detailed design of our Conditional Guidance Adapter.

verse this noise addition. The process begins with the diffusion phase, where an initial image $\boldsymbol{x}_0$ evolves into Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0,1)$ through T steps. Each step of this transformation is governed by:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right)=\mathcal{N}\left(\boldsymbol{x}_t ; \sqrt{1-\beta_t} \boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right), \quad (1)$$

where $\boldsymbol{x}_t$ is the image with noise at step t, β_t is a fixed scaling factor, and $\mathcal{N}$ denotes the Gaussian distribution. Introducing $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$ simplifies Eq.1 to:

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

During the inference stage's reverse process, DMs initiate with a random Gaussian noise map $\boldsymbol{x}_T$ and progressively denoise it to reach the refined output $\boldsymbol{x}_0$.

$$p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)=\mathcal{N}\left(\boldsymbol{x}_{t-1} ; \boldsymbol{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right), \sigma_t^2 \boldsymbol{I}\right), \quad (3)$$

where the mean $\boldsymbol{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \boldsymbol{\epsilon}\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right)$ and the variance $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

3.2. Overview of DTP-IR Framework

Our aim is to create a unified framework for image restoration tasks based on our *Diffusion Texture Priors*. As shown in Figure 3, our DTPM framework comprises two stages: In Stage 1, we train a conditional diffusion model with semantic latent feature constraints using a large dataset of high-resolution, high-quality natural images. During Stage 2, we incorporate conditional guidance adapters into the conditional diffusion model to facilitate the model's adaptation to downstream image restoration tasks.

3.3. Stage I: Learning Diffusion Texture Priors

In this section, we will first briefly introduce the core motivation of our Diffusion Texture Priors, followed by the basic structure information of the diffusion model and the semantic code. Finally, we will discuss the training objective and training data.

Motivation. Recent studies [52, 64, 74] have highlighted the pivotal role of texture quality in shaping subjective visual perception in the realm of image restoration. Traditional methodologies, however, encounter substantial difficulties in accurately restoring fine-grained textural details, which are crucial for realistic image reconstruction. While diffusion models have demonstrated excellence in generating lifelike images, their direct application to image restoration tasks has been less than ideal, yielding mediocre results. Our approach seeks to remedy this by integrating

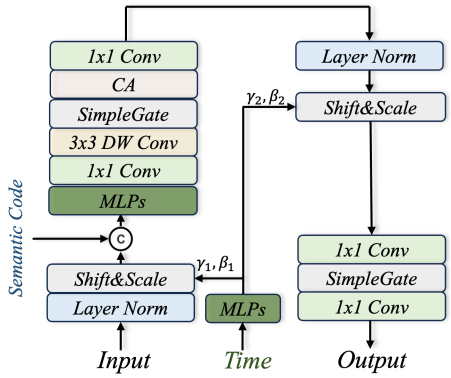

Figure 4. The architecture of our fundamental denoising block incorporates "Channel Attention" (CA) and "SimpleGate," a straightforward feature gating mechanism introduced by Chen *et al.*[8]. The current timestep t is injected into the block with Adaptive Layer Normalization [3] (AdaLN), *i.e.*, $AdaLN(f_{in} = \gamma LayerNorm(f_{in}) + \beta$, where f_{in} is the input features, γ and β are converted from the timestep embedding. The "Shift&Scale" block means shift and scale features by generated γ, β.

texture-specific priors into the diffusion model. This integration enables the diffusion model to specifically concentrate on reconstructing textures, as opposed to merely focusing on degradation elimination. ***Such a strategy leverages the diffusion model's inherent strengths and mitigates its weaknesses, allowing us to exploit a vast repository of high-quality natural images for prior knowledge.*** This method not only significantly enhances the model's generalization capabilities but also markedly improves the efficiency of the sampling process.

Basic Denoising Block. We have engineered a denoising block based on NAFNet [8] block, designed to enhance model adaptability for a broad spectrum of natural image content. This block aims to balance peak model performance with reduced computational demands and easy-to-handle varying input resolutions. As delineated in Figure 4, our denoising block capitalizes on depth-wise convolution as its cornerstone, complemented by several 1×1 convolution layers and LayerNorm [3], to facilitate efficient denoising. To incorporate the time condition, we utilize MLPs for transforming the time embedding into channel-specific scale and shift parameters. Additionally, to maintain the semantic consistency between x_0 and y_s, we integrate an MLP layer for semantic condition, which amalgamates semantic code as a pivotal conditioning factor.

Semantic code. To enable the diffusion model to produce realistically textured details, we utilize a pre-trained ResNet-18 [21] for extracting semantic embeddings as a global semantic code, which is then integrated into each denoising block as shown in Figure 4. *This approach of employing semantic code imposes robust semantic constraints, rather than solely depending on adjacent context information, allowing our model to generate textures in alignment with the contextual semantics.* Consequently, this ensures high-fidelity results of our diffusion model.

Training objective. To prioritize texture reconstruction over content of the image, our diffusion model is designed to capture the residual distribution of y_s. The training objective for Stage I is thus defined as: $L_{\text{StageI}}(\theta) =$

$$\mathbb{E}\left\|\boldsymbol{\epsilon} - f_\theta\left(\sqrt{\gamma}\left(\underbrace{\boldsymbol{x}_g - \boldsymbol{y}_s}_{\text{residual}}\right) + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \boldsymbol{y_s}, \gamma\right)\right\|, \quad (4)$$

here, $\epsilon \sim \mathcal{N}(0, I)$ represents the noise vector, and f_θ is our diffusion model. The term $\gamma \sim p(\gamma)$ denotes a distribution associated with the noise schedule. In this context, x_g symbolizes the high-quality input image, whereas y_s denotes its corresponding smoothed image. y_s is obtained by utilizing an edge-preserving smoothing method [63] to eliminate a majority of the texture while retaining the contextual content.

Training data for Stage I. Inspired by previous codebook prior-based image restoration studies [61, 64, 74], our training stage I employs high-quality images from the widely-recognized DIV2K [2] and Flickr2K [35] datasets. Each image is cropped into multiple patches of 512×512 resolution, followed by an edge-preserved smooth method [63] to diminish textural details. This process results in a training dataset comprising 55,558 paired instances (x_g, y_s).

3.4. Stage II: Finetuning Diffusion Texture Priors Model with Adapters

This section describes the initial predictor and the Conditional Guidance Adapter, along with the rationale behind their integration. We also detail the architecture of the adapters. Finally, the training objectives are outlined.

Initial Predictor. In the pursuit of augmenting the diffusion model's adaptability for downstream tasks in its second stage (see Figure 3), we draw from the DvSR, integrating a compact and lightweight U-Net to generate an initial output. This initial step efficiently seizes deterministic components of the ultimate restored image, crucial for capturing essential structural information. Moreover, this initial output serves as a pivotal guiding element for the diffusion model, forming an effective complement with our Diffusion Texture Priors Model.

Table 1. **Image Desnowing.**

Method	Snow100K-S [40]			Snow100K-L [40]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SPANet [58]	29.92	0.8260	0.104	23.70	0.7930	0.210
JSTASR [12]	31.40	0.9012	0.059	25.32	0.8076	0.152
RESCAN [34]	31.51	0.9032	0.054	26.08	0.8108	0.106
DesnowNet [40]	32.33	0.9500	0.070	27.17	0.8983	0.101
DDMSNet [72]	34.34	0.9445	0.058	28.85	0.8772	0.098
MPRNet [71]	34.97	0.9457	0.049	29.76	0.8949	0.091
NAFNet [8]	34.79	0.9497	0.051	30.06	0.9017	0.086
Restormer [70]	35.03	0.9487	0.047	30.52	0.9092	0.083
SnowDiff$_{64}$[45]	36.59	0.9626	0.035	30.43	0.9145	0.069
SnowrDiff$_{128}$[45]	36.09	0.9545	0.038	30.28	0.9000	0.067
DTPM-4 Step	**37.01**	**0.9663**	0.034	**30.92**	**0.9174**	0.063
DTPM-10 Step	36.44	0.9521	0.033	30.32	0.9024	0.055
DTPM-50 Step	36.19	0.9483	**0.029**	30.02	0.8901	**0.053**

Table 2. **Image Deraining & Dehazing.**

Method	Outdoor-Rain [33]		
	PSNR ↑	SSIM ↑	LPIPS ↓
CycleGAN [75]	17.62	0.6560	-
pix2pix [23]	19.09	0.7100	-
HRGAN [33]	21.56	0.8550	0.154
PCNet [26]	26.19	0.9015	0.132
MPRNet [71]	28.03	0.9192	0.089
NAFNet [8]	29.59	0.9027	0.085
Restormer [70]	29.97	0.9215	0.074
RainHazeDiff$_{64}$[45]	28.38	0.9320	0.067
RainHazeDiff$_{128}$[45]	26.84	0.9152	0.071
DTPM-4 Step	**30.99**	**0.934**	0.0635
DTPM-10 Step	30.92	0.932	0.0617
DTPM-50 Step	30.48	0.921	**0.0540**

Table 3. **Removing Raindrops.**

Method	RainDrop [48]		
	PSNR ↑	SSIM ↑	LPIPS ↓
pix2pix [23]	28.02	0.8547	–
DuRN [38]	31.24	0.9259	–
RaindropAttn [50]	31.44	0.9263	0.068
AttentiveGAN [48]	31.59	0.9170	0.055
CCN [49]	31.34	0.9286	0.066
IDT [62]	31.87	0.9313	0.059
RainDropDiff$_{64}$[45]	32.29	0.9422	0.058
RainDropDiff$_{128}$[45]	32.43	0.9334	0.058
DTPM-4 Step	**32.72**	**0.9440**	0.0577
DTPM-10 Step	31.84	0.9370	0.0477
DTPM-50 Step	31.44	0.9320	**0.0439**

Conditional Guidance Adapter. DPTM encapsulates rich and diverse prior knowledge. We hope that DPTM can fully utilize this knowledge in adapting to downstream image restoration tasks, rather than forgetting it. Therefore, we freeze most of the parameters in DPTM and introduce a flexible and parameter-efficient adaptation mechanism through the Conditional Guidance Adapter as shown in Figure 3 (b) and (c). The design of the Conditional Guidance Adapter brings two benefits: (i) *Parameter-efficient learning.* Since most of the parameters of DPTM have been sufficiently trained, our model will converge faster and better on downstream tasks after introducing the adapter. (ii) *Introducing additional conditions.* Introduction of coarse structural information output by the initial predictor, providing the model with more flexible conditional information, thereby producing texture details that better align with expectations. Our Conditional Guidance Adapter could be formally written as:

$$\text{CG-Adapter}(f_{in}, y_\epsilon) = f_{in} + \text{M}_{out}(\text{M}_{in}(f_{in}) + \text{Conv}(y_\epsilon)), \tag{5}$$

where M denotes an MLP layer to scale the dimension c of the input feature f_{in}^c to dimension $r \times c$, and y_ϵ denotes the output of initial predictor ϵ_θ, thus $y_\epsilon = \epsilon_\theta(y_d)$.

Training objective. The training objective for Stage II is defined as: $L_{\text{StageII}}(\theta) =$

$$\mathbb{E}\left\|\boldsymbol{\epsilon} - f_\theta\left(\sqrt{\gamma}\left(\boldsymbol{y}_g - \epsilon_\theta(\boldsymbol{y}_d)\right) + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \boldsymbol{y_d}, \gamma\right)\right\|. \tag{6}$$

where ϵ_θ represents the initial predictor, illustrated in Figure 3. It is important to note that ϵ_θ does not necessitate additional supervisory loss or pre-training. This is because the gradient from the aforementioned loss propagates through f_θ and subsequently influences ϵ_θ.

4. Experiments

4.1. Settings

Our method is assessed on **five specific image restoration tasks**: image motion deblurring, single-image defocus deblurring, image deraining&dehazing, image desnowing, and image raindrop removal. In the initial stage, we develop a comprehensive Diffusion Texture Prior model, and in the subsequent stage, we fine-tune individual full models tailored to each task.

Datasets&Metrics. The GoPro dataset [43], as recommended by DvSR [60], is utilized for training and testing in the domain of image motion deblurring. Adhering to Restormer's approach [70], the DPDD dataset [1] is employed for single-image defocus deblurring. For image desnowing, we align with WeatherDiff [45] using the Snow100K dataset [40]. The Outdoor-Rain dataset [33] is used for image deraining&dehazing, following WeatherDiff [45]. Lastly, for single image raindrop removal, we adopt the practices from IDT [62] and WeatherDiff [45], employing the Raindrop dataset [48] for training and testing. We utilize PSNR, SSIM, and MAE as metrics for distortion-based image quality assessment, complemented by LPIPS [73] and FID for evaluating perceptual image quality. Adhering to the methodology outlined in Whang et al. [60], we extract image patches and compute FID at the patch level to ensure more consistent and reliable evaluation results.

Training Details We implemented our DTPM using the Pytorch framework, leveraging four NVIDIA RTX 4090 GPUs. Training encompasses two stages: Stage I with 800K iterations and Stage II with 600K iterations. We employed the Adam optimizer, setting momentum values at 0.9 and 0.999. The initial learning rate was established at 1.5×10^{-4}, employing a cosine annealing strategy for gradual learning rate reduction. The diffusion process is

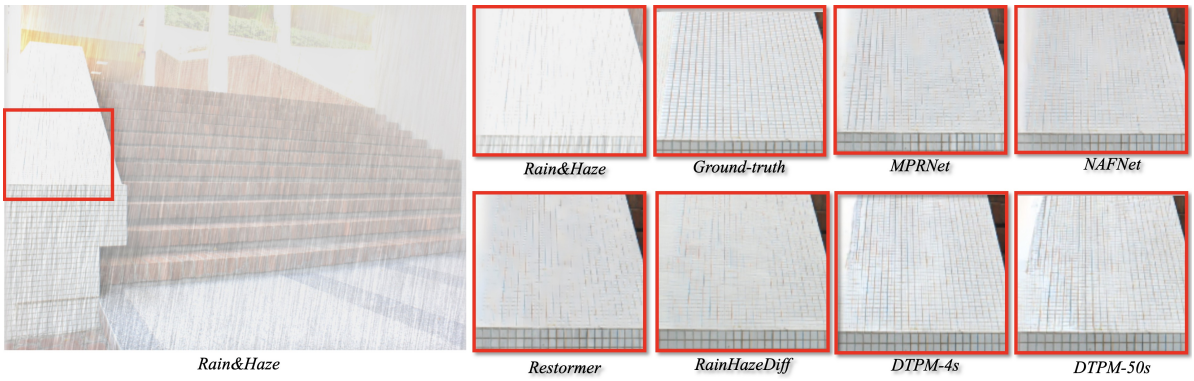

Figure 5. Image deraining&dehazing visual comparison on the Ourdoor-Rain dataset [48]. Our DTPM method generates rain-free and haze-free image with better details and without artifacts. The "-4s" and "-50s" denotes our DTPM method with a DDIM [55] sampling schedule of 4 steps and 50 steps.

Figure 6. Single-image motion deblurring comparison on the GoPro Dataset [43]. Compared to the other methods, our DTPM more effectively removes blur and preserving better structural information. The "-4s" and "-50s" denotes our DTPM method with a DDIM sampling schedule of 4 steps and 50 steps.

structured over 1,000 steps (denoted as T), incorporating a noise schedule $\beta(t)$ that linearly escalates from 0.0001 to 0.02 throughout the training. Input data consistently utilized 256×256 cropped patches. Data augmentation techniques, including horizontal flipping and random image rotation at $45°$ and $90°$, were applied during training. We utilize Denoising Diffusion Implicit Models (DDIM) [55] to implement our diffusion model, thereby significantly enhancing our sampling speed.

4.2. Main Results

Comparative Analysis of Deraining&Dehazing. As shown in Table 1 in outdoor hazy and rainy image restoration, DTPM's 4 Step variant outperforms RainHazeDiff's 25-step process, achieving higher PSNR (30.99), SSIM (0.934), and lower LPIPS (0.0635). DTPM demonstrates superior perceptual quality enhancement, as indicated by its LPIPS and PSNR metrics. The DTPM-10 and DTPM-50 Step versions further reduce LPIPS to 0.0617 and 0.0540, respectively, while maintaining high SSIM, establishing DTPM's methodological advantage and benchmark in high-fidelity restoration.

Comparative Analysis of Single-image Defocus Deblurring. DTPM markedly outperforms other methods in defocus deblurring on the DPDD dataset as shown in Table 4, with its 4 Step iteration achieving high LPIPS (0.153) and SSIM (0.823) scores, thus maintaining image structural integrity. The DTPM-50 Step model further enhances perceived quality, achieving an LPIPS of 0.139, closely resembling ground truth. Visual evidence in Figure 7 corroborates DTPM's superior detail recovery and structural fidelity.

Concise Analysis of DTPM in Single-image Motion Deblurring. As shown in Table 5,on the GoPro dataset, the DTPM 4 Step variant achieves a PSNR of 32.09, SSIM of 0.932, LPIPS of 0.084, and FID of 10.02, demonstrating a balance in distortion reduction and perceptual quality. The 10 Step version improves these metrics, with an LPIPS of 0.081 and a leading FID of 8.52, despite slight decreases in PSNR and SSIM. The DTPM-50 Step variant, achieving the best perceptual outcomes with the lowest LPIPS (0.073) and FID (7.69), slightly reduces PSNR to 31.11 and SSIM to 0.919, illustrating our method's ability to the best perceptual quality.

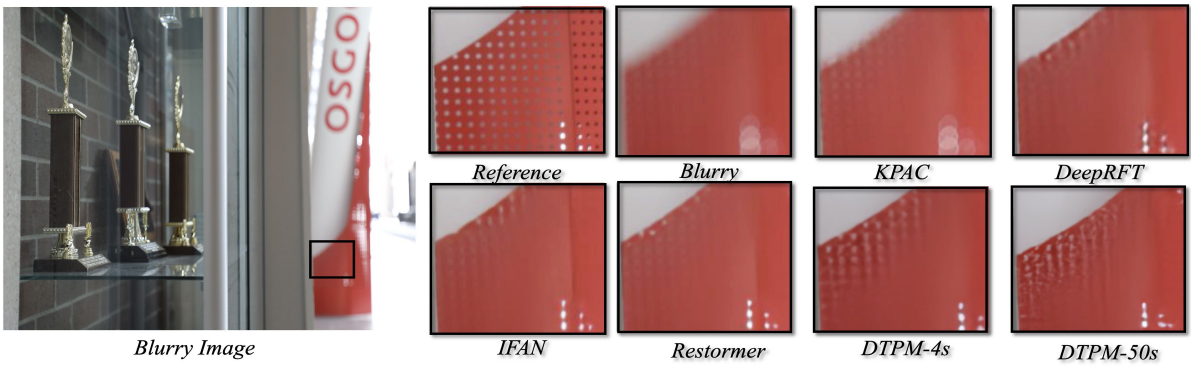

Figure 7. Single-image defocus deblurring results on the DPDD dataset [1]. The "-4s" and "-50s" denotes our DTPM method with a DDIM sampling schedule of 4 steps and 50 steps. Compared to the other approaches, our DTPM generate better structures and details.

Table 4. **Single-image Defocus Deblurring** comparisons on the DPDD testset [1].

Method	Distortion			Perceptual
	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓
EBDB [28]	23.45	0.683	0.049	0.336
DMENet [31]	23.41	0.714	0.051	0.349
JNB [53]	23.84	0.715	0.048	0.315
DPDNet [1]	24.34	0.747	0.044	0.277
KPAC [54]	25.22	0.774	0.040	0.227
IFAN [32]	25.37	0.789	0.039	0.217
Restormer [70]	25.98	0.811	0.038	0.178
DTPM-4 Step	**25.98**	**0.823**	**0.038**	0.153
DTPM-10 Step	25.76	0.815	0.039	0.140
DTPM-50 Step	25.45	0.803	0.040	**0.139**

Table 5. **Single-image Motion Deblurring** comparisons on the GoPro testset [43].

Method	Distortion		Perceptual	
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
MPRNet [71]	32.66	0.959	0.089	20.18
MIMO-UNet+ [13]	32.45	0.957	0.091	18.05
SAPHNet [56]	31.89	0.953	0.101	19.06
Restormer [70]	33.20	0.963	0.084	19.33
DvSR [60]	30.66	0.941	0.084	12.20
DTPM-4 Step	**32.09**	**0.932**	0.084	10.02
DTPM-10 Step	31.82	0.929	0.077	7.52
DTPM-50 Step	31.11	0.919	**0.061**	**6.61**

Comparative Analysis of Raindrop Removal. As illustrated in Table 2, we can observe that various versions of DTPM have demonstrated outstanding performance in the task of image raindrop removal. In comparison to RainDropDiff, our approach has achieved superior performance, with the LPIPS metric significantly outperforming that of RainDropDiff.

Comparative Analysis of Snow Removal. As shown in Table 1, we can observe that various versions of DTPM have demonstrated excellent performance in the task of image desnowing. In comparison to SnowDiff, our approach has achieved superior LPIPS metrics.

5. Discussion and Analysis

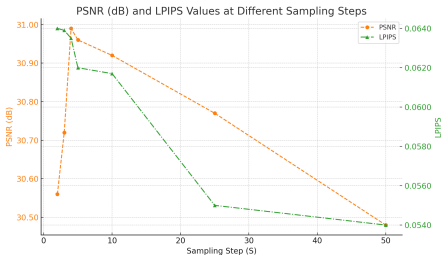

Figure 8. Trade-off between PSNR and LPIPS in DTPM: This plot illustrates the inverse relationship between PSNR (pixel-wise distortion) and LPIPS (perceptual similarity) across various sampling steps, highlighting the equilibrium achieved between distortion and perceptual quality in our method.

For the analysis of various aspects of our model, we perform our experiments on the image deraining& dehazing dataset as it's a more chanllenging settings than others.

5.1. Sampling Steps and Distortion-Perceptual Trade-off

In Figure 8, we evaluate the results of our DTPM through the lens of two critical image quality metrics: PSNR and LPIPS[1]. As the sampling steps increase, the PSNR values initially experience a precipitous drop, decreasing from just above 31 dB to below 30.70 dB by the 10th step. This substantial early decline indicates that sampling steps in the early phase degrade image fidelity. Beyond this point, the PSNR decline slows, entering a phase of more gradual descent before plateauing after the 30th step, suggesting that the method begins to stabilize in terms of distortion loss. In

[1]The PSNR, measured in decibels (dB), is a traditional metric for assessing the fidelity of the reconstructed images against an original high-quality image. The LPIPS metric, on the other hand, offers an assessment of perceptual similarity, which correlates more closely with human visual perception.

Figure 9. The first row in the figure shows the images, while the second row shows zoomed-in details. From left to right, they are: **(a) input image, (b) the output of the initial predictor, (c) the final result of our method, (d) the ground-truth.**

contrast, the LPIPS score improves markedly with each incremental sampling step, indicating that the perceived quality of the images is becoming more aligned with human visual perception. The steepness of the LPIPS improvement mirrors the PSNR decline, reflecting a pronounced trade-off between fidelity and perceptual quality up to the 10 step. After this inflection point, although the PSNR values level off, suggesting a saturation of fidelity loss, the LPIPS values continue to improve but at a diminished rate, plateauing around the 40 step. *The intersecting trends of these metrics underscore a distinct trade-off: optimizing for perceptual likeness through more sampling steps comes at the expense of traditional fidelity measures.* However, the stabilization of both metrics after a certain number of sampling steps suggests a convergence towards an equilibrium between the two qualities.

5.2. Output of the Initial Predictor

Given the significance of the initial predictor as a key component of our methodology, we have undertaken detailed explorations and analyses of it. Despite the absence of an explicit loss constraint in our initial predictor, it achieves results that closely resemble a clean background, as evidenced in Figure 9. The initial predictor, though not detailed in its finer aspects, effectively produces a reasonably clear image. *This provides practical assurance for incorporating the output of the initial predictor as one of the enhanced, additional conditions for diffusion.*

5.3. More Ablation Studies

In this section, we conduct ablation studies to analyze the impact of different design choices in our DTPM framework.
w/o Stage I Training. This variant, absent the initial training phase, exhibits a decrease in both image reconstruction quality and perceptual similarity relative to the complete model. This underscores the significance of the early training stage in establishing baseline performance.
w/o Initial Predictor. Removing the Initial Predictor results in a notable decline in both objective image quality and perceptual accuracy. This component's role is evidently crucial for enhancing the clarity and fidelity of the images.
w/o Semantic Code. The removal of the Semantic Code somewhat improves both image quality and perceptual similarity. This could suggest that the Semantic Code potentially adding informative features.
Fine-tune Enc. w/o Adapters. Fine-tuning just the encoder of DTPM without adapters leads to a slight reduction in performance. This indicates that while the encoder is a robust component, the adapters contribute to optimizing its function.
Fine-tune Dec. w/o Adapters. Solely fine-tuning the decoder of DTPM without adapters also reduces model efficacy. The adapters appear more integral for the decoder, possibly enabling more nuanced adjustments that refine results.
Fine-tune Enc.&Dec. w/o Adapters. Fine-tuning both the encoder and decoder without adapters does not match the performance of the complete model. This suggests a complex interplay between these components that adapters help to fine-tune effectively.

Table 6. The ablation studies on Image Deraining&Dehazing dataset [48].

Model	PSNR ↑	LPIPS ↓
w/o Stage I Training	28.90	0.074
w/o Initial Predictor	26.91	0.079
w/o Semantic Code	29.79	0.0653
Fine-tune Enc. w/o Adapters	29.20	0.0667
Fine-tune Dec. w/o Adapters	28.91	0.0691
Fine-tune Enc.&Dec. w/o Adapters	29.13	0.0712
Ours(DTPM-4S)	**30.99**	**0.0635**

6. Conclusion and Limitations

This study presents a comprehensive conditional framework for image restoration, integrating Diffusion Texture Priors for producing high-quality images with promising details. While achieving significant performance, it identifies potential improvements: accelerating sampling via advanced techniques and optimizing the model size.

Acknowledgment. This work is supported by the InnoHK funding launched by Innovation and Technology Commission, Hong Kong SAR, Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0671), and a grant of Innovation and Technology Fund under Guangdong-Hong Kong Technology Cooperation Funding Scheme (ITF-TCFS) (project no. GHP/051/20GD).

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *ECCV*, 2020. 5, 7

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 4

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 1, 2

[5] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 1

[6] Shidong Cao, Wenhao Chai, Shengyu Hao, and Gaoang Wang. Image reference-guided fashion design with structure-aware transfer by diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3524–3528, 2023. 2

[7] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE Transactions on Multimedia*, 2023. 2

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022. 1, 4, 5

[9] Sixiang Chen, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, and Lei Zhu. Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13106–13117, 2023. 2

[10] Sixiang Chen, Tian Ye, Yun Liu, Taodong Liao, Yi Ye, and Erkang Chen. Msp-former: Multi-scale projection transformer for single image desnowing. *arXiv preprint arXiv:2207.05621*, 2022.

[11] Sixiang Chen, Tian Ye, Jun Shi, Yun Liu, JingXia Jiang, Erkang Chen, and Peng Chen. Dehrformer: Real-time transformer for depth estimation and haze removal from varicolored haze scenes. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[12] Wei-Ting Chen, Hao-Yu Fang, Jian-Jiun Ding, Cheng-Che Tsai, and Sy-Yen Kuo. Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In *European Conference on Computer Vision*, pages 754–770. Springer, 2020. 5

[13] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 7

[14] Jie Deng, Wenhao Chai, Jianshu Guo, Qixuan Huang, Wenhao Hu, Jenq-Neng Hwang, and Gaoang Wang. Citygen: Infinite and controllable 3d city layout generation. *arXiv preprint arXiv:2312.01508*, 2023. 1

[15] David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proceedings of the IEEE international conference on computer vision*, pages 633–640, 2013. 2

[16] Jianshu Guo, Wenhao Chai, Jie Deng, Hsiang-Wei Huang, Tian Ye, Yichen Xu, Jiawei Zhang, Hwang Jenq-Neng, and Gaoang Wang. Versat2i: Improving text-to-image models with versatile reward. *arXiv preprint arXiv:2403.18493*, 2024. 1

[17] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 1, 2

[18] Chunming He, Chengyu Fang, Yulun Zhang, Kai Li, Longxiang Tang, Chenyu You, Fengyang Xiao, Zhenhua Guo, and Xiu Li. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. *arXiv preprint arXiv:2311.11638*, 2023. 2

[19] Chunming He, Kai Li, Guoxia Xu, Jiangpeng Yan, Longxiang Tang, Yulun Zhang, Yaowei Wang, and Xiu Li. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[20] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 1, 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 5

[24] Da-Wei Jaw, Shih-Chia Huang, and Sy-Yen Kuo. Desnowgan: An efficient single image snow removal framework using cross-resolution lateral connection and gans. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1342–1350, 2020. 2

[25] Jingxia Jiang, Tian Ye, Jinbin Bai, Sixiang Chen, Wenhao Chai, Shi Jun, Yun Liu, and Erkang Chen. Five a^{+} network: You only need 9k parameters for underwater image enhancement. *arXiv preprint arXiv:2305.08824*, 2023. 2

[26] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Zheng Wang, Xiao Wang, Junjun Jiang, and Chia-Wen Lin. Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining. *IEEE Transactions on Image Processing*, 30:7404–7418, 2021. 5

[27] Yeying Jin, Wenhan Yang, Wei Ye, Yuan Yuan, and Robby T

Tan. Shadowdiffusion: Diffusion-based shadow removal using classifier-driven attention and structure preservation. *arXiv preprint arXiv:2211.08089*, 2022. 2

[28] Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *TIP*, 2017. 7

[29] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 2

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[31] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *CVPR*, 2019. 7

[32] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, 2021. 7

[33] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1633–1642, 2019. 5

[34] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018. 5

[35] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 4

[36] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 1

[37] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7314–7323, 2019. 1

[38] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 5

[39] Yun Liu, Zhongsheng Yan, Aimin Wu, Tian Ye, and Yuche Li. Nighttime image dehazing based on variational decomposition model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 640–649, 2022. 2

[40] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018. 5

[41] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023. 2

[42] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 1

[43] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 1, 5, 6, 7

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 2

[45] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 5

[46] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[47] Yan-Tsung Peng, Keming Cao, and Pamela C Cosman. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, 27(6):2856–2868, 2018. 1, 2

[48] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2482–2491, 2018. 2, 5, 6, 8

[49] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9147–9156, 2021. 5

[50] Yuhui Quan, Shijie Deng, Yixin Chen, and Hui Ji. Deep learning for seeing through window with raindrops. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2463–2471, 2019. 5

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[52] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2, 3

[53] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *CVPR*, 2015. 7

[54] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *ICCV*, 2021. 7

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*

arXiv:2010.02502, 2020. 1, 6

[56] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3606–3615, 2020. 7

[57] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023. 1

[58] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12270–12279, 2019. 5

[59] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 2

[60] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 2, 5, 7

[61] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22282–22291, 2023. 2, 4

[62] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5

[63] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012. 4

[64] Tian Ye, Sixiang Chen, Jinbin Bai, Jun Shi, Chenghao Xue, Jingxia Jiang, Junjie Yin, Erkang Chen, and Yun Liu. Adverse weather removal with codebook priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12653–12664, 2023. 2, 3, 4

[65] Tian Ye, Sixiang Chen, Yun Liu, Wenhao Chai, Jinbin Bai, Wenbin Zou, Yunchen Zhang, Mingchao Jiang, Erkang Chen, and Chenghao Xue. Sequential affinity learning for video restoration. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4147–4156, 2023. 2

[66] Tian Ye, Sixiang Chen, Yun Liu, Yi Ye, Jinbin Bai, and Erkang Chen. Towards real-time high-definition image snow removal: Efficient pyramid network with asymmetrical encoder-decoder architecture. In *Proceedings of the Asian Conference on Computer Vision*, pages 366–381, 2022. 1, 2

[67] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *European Conference on Computer Vision*, pages 130–145. Springer, 2022. 1, 2

[68] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 1

[69] Hu Yu, Jie Huang, Kaiwen Zheng, Man Zhou, and Feng Zhao. High-quality image dehazing with diffusion model. *arXiv preprint arXiv:2308.11949*, 2023. 1

[70] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1, 2, 5, 7

[71] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 1, 2, 5, 7

[72] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing*, 30:7419–7431, 2021. 5

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[74] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 2, 3, 4

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 5

[76] Wenbin Zou, Hongxia Gao, Tian Ye, Liang Chen, Weipeng Yang, Shasha Huang, Hongsheng Chen, and Sixiang Chen. Vqcnir: Clearer night image restoration with vector-quantized codebook. *arXiv preprint arXiv:2312.08606*, 2023. 2

[77] Wenbin Zou, Tian Ye, Weixin Zheng, Yunchen Zhang, Liang Chen, and Yi Wu. Self-calibrated efficient transformer for lightweight super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 930–939, 2022. 2