This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Visual-Instructed Degradation Diffusion for All-in-One Image Restoration

Wenyang Luo[1,2*], Haina Qin[1,2*♡], Zewen Chen[1,2*], Libin Wang[3†],
Dandan Zheng[3], Yuming Li[3], Yufan Liu[1,2], Bing Li[1,2] , Weiming Hu[1,2]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), CASIA;
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences; [3]Ant Group;
{luowenyang2020@,qinhaina2020@,bli@nlpr.,wmhu@nlpr.}ia.ac.cn lbin.wlb@antgroup.com

Abstract

Image restoration tasks like deblurring, denoising, and dehazing usually need distinct models for each degradation type, restricting their generalization in real-world scenarios with mixed or unknown degradations. In this work, we propose ***Defusion****, a novel all-in-one image restoration framework that utilizes visual instruction-guided degradation diffusion. Unlike existing methods that rely on task-specific models or ambiguous text-based priors, Defusion constructs explicit* ***visual instructions*** *that align with the visual degradation patterns. These instructions are grounded by applying degradations to standardized visual elements, capturing intrinsic degradation features while agnostic to image semantics. Defusion then uses these visual instructions to guide a diffusion-based model that operates directly in the degradation space, where it reconstructs high-quality images by denoising the degradation effects with enhanced stability and generalizability. Comprehensive experiments demonstrate that Defusion outperforms state-of-the-art methods across diverse image restoration tasks, including complex and real-world degradations.*

1. Introduction

Image restoration is a fundamental task in computer vision, aiming to recover high-quality images from their degraded counterparts. This encompasses various challenges, including denoising [8, 95, 121], deblurring [110, 111, 122], dehazing [64, 109], deraining [16, 17, 42], JPEG artifact removal [25, 41], etc. Traditional deep learning-based approaches have achieved remarkable progress by developing task-specific models that excel in individual restoration tasks. However, these approaches often struggle with generalization, requiring separate models for each degradation type, which limits their practical applicability in real-world scenarios with complex degradations.

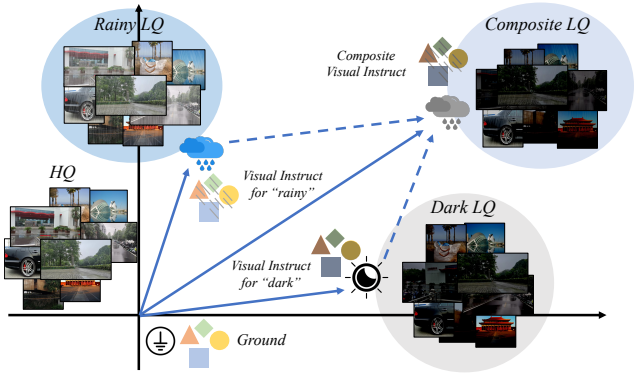


Figure 1. Visual instructions speak for visual degradations. Visual instructions are constructed by applying the artifacts of image degradations upon the composition of basic image elements (dubbed visual grounds). They faithfully demonstrate the visual effects of degradations and their compositions.

To address this limitation, all-in-one image restoration methods handling various degradations in a single model have gained significant attention [4, 59, 65, 84, 86, 143]. They use implicit or explicit prior knowledge to identify degradation patterns and guide restoration. However, existing methods face challenges in deriving effective priors that align with the visual structures of degraded images.

Implicit priors are extracted from input low-quality (LQ) images by task-specific sub-networks [14, 84] or learnable prompts [4, 59, 86, 143]; they are essentially equivalent to increasing the parameter counts, limiting their generalizability to complex degradations [43]. Explicit priors, typically text descriptions processed by language models [19, 131], suffer from weak alignment with low-level visual structures [66, 124]. Text descriptions are typically coarse or agnostic to degradations [13] due to inherent language ambiguity and a lossy language-vision bridge. Thus, the text descriptions may collapse into category labels, crippling the benefit of the added complexity.

Given the above limitations, we propose to *let visual*

*Equal contribution. † Project lead. Project Page.
♡Work done during internship at Ant Group.

speak for visual and guide the image restoration model with **visual instructions** as explicit priors. Our visual instructions are constructed to demonstrate the artifacts of image degradations as visual guidance. This is achieved by applying the image corruption process (e.g. raining or noising) of the LQ images upon the visual grounds consisting of basic visual elements such as random textures, basic shapes, standard colors, etc, as shown in Fig. 1. The resulting visual instructions encode the exact visual effects (e.g. rain lines or noisy "dots") of the degradation with a minimal loss of information. The primary benefit is that the resulting visual instructions are inherently aligned with visual degradation patterns. For example, the combined visual instruction in Fig. 1 can naturally show similar distortions as the composite LQ images. This makes the visual instructions suitable for low-level vision tasks.

We propose a unified diffusion-based image restoration framework, **De**gradation Dif**fusion**, which is guided by our visual instructions to denoise in the degradation space. The diffusion model provides generative priors for the restored images, conditioned on the LQ images through an Image Restoration Bridge (IRB) to preserve the structural priors of the LQ images and maintain visual consistency between LQ and restored images. Visual instructions, which capture the prominent intrinsic artifacts that degradations induce on textures, colors, lightness, and other image elements, control the model to perform restoration tasks. Specifically, visual instructions are "tokenized" and hint the model of the degradations in the LQ images via a Visual Instruct Adapter (VIA). Moreover, Defusion operates within the degradation space defined as the space of transformations between LQ and HQ images, which aligns with the visual instructions describing the degradation. The restoration model is jointly trained on multiple tasks. During inference, it follows visual instructions to handle various degradation types.

Our contributions are summarized as follows:

1. We introduce a novel diffusion-based all-in-one image restoration framework, dubbed Defusion, guided by visual instructions to denoise in the degradation space.
2. Our visual instructions visually align with the image degradations and retain their visual compositional properties, giving them advantages in low-level vision tasks.
3. We conducted comprehensive experiments on various image restoration tasks under different scenarios including complex and unseen degradations, and the results demonstrate that Defusion consistently outperforms state-of-the-art methods.

2. Related Work

2.1. All-in-One Image Restoration

Image restoration methods mainly target predefined image degradation [15, 28, 31, 44, 52, 53, 57, 60, 134, 148] such as motion blurring or rainy images. Such methods require developing and deploying separate models for each kind of degradation. Recently, several all-in-one approaches [4, 19, 34, 43, 51, 59, 65, 74, 84, 86, 88, 130, 131, 143, 145, 147] have been proposed to tackle multiple degradations with one unified model. All-in-one approaches have the benefit that the model can learn the common knowledge from mixed degradations and possibly generalize to composite or unseen degradation.

Most all-in-one approaches [4, 34, 51, 59, 65, 74, 84, 86, 88, 130, 143, 145, 147] adopt auxiliary modules or learnable prompts to detect degradation types for the restoration model. These designs are often equivalent to increasing model parameter count, which hinders their generalizability to complex degradations.

SUPIR [131], InstructIR [19], and AutoDIR [43] guide image restoration with text-based instructions and utilize the concepts the pretrained language models learned to generalize to multiple degradations. However, language is essentially ambiguous and weakly aligned with visual features. For example, "enhance the image" can either mean "increase lightness" or "denoising"; "underwater image enhancement" can be visually decomposed into "lightness enhancement" and "deblurring", which is not obvious from language.

In contrast, we propose accurately guiding image restoration with visual instructions, which closely align with the underlying visual degradations. Our visual instructions directly reveal distorted images' visual artifacts. The restoration model can learn the individual distortion patterns and their combinations during training and generalize to new degradations with the hints of our visual instructions.

2.2. Diffusion Models for Image Restoration

Image restoration can be formulated as an image-to-image translation task, with low-quality (LQ) images as input and high-quality (HQ) images as output. Diffusion-based generative models (DMs) [36, 106, 108] have risen as attractive solutions due to their remarkable ability to produce natural and visually appealing HQ images [22, 23, 45, 73, 97].

Many works [4, 43, 59, 65, 73, 125, 131, 144] adopt the latent diffusion architecture following Stable Diffusion [97], where a separate autoencoder (VAE) [47] extracts latent codes from the LQ images to denoise. However, the VAE discards most low-level details, which harms image restoration.

Other works operate on pixels [45, 62, 72] or frequency domain [146]. Among them, some works propose dedicated diffusion processes to model image degradations, where the HQ images are corrupted given the LQ images [45] or shifts between HQ and LQ images [62, 63, 72, 105, 132]. These complicated approaches usually make approximate assumptions about HQ/LQ images or their shifts, leading

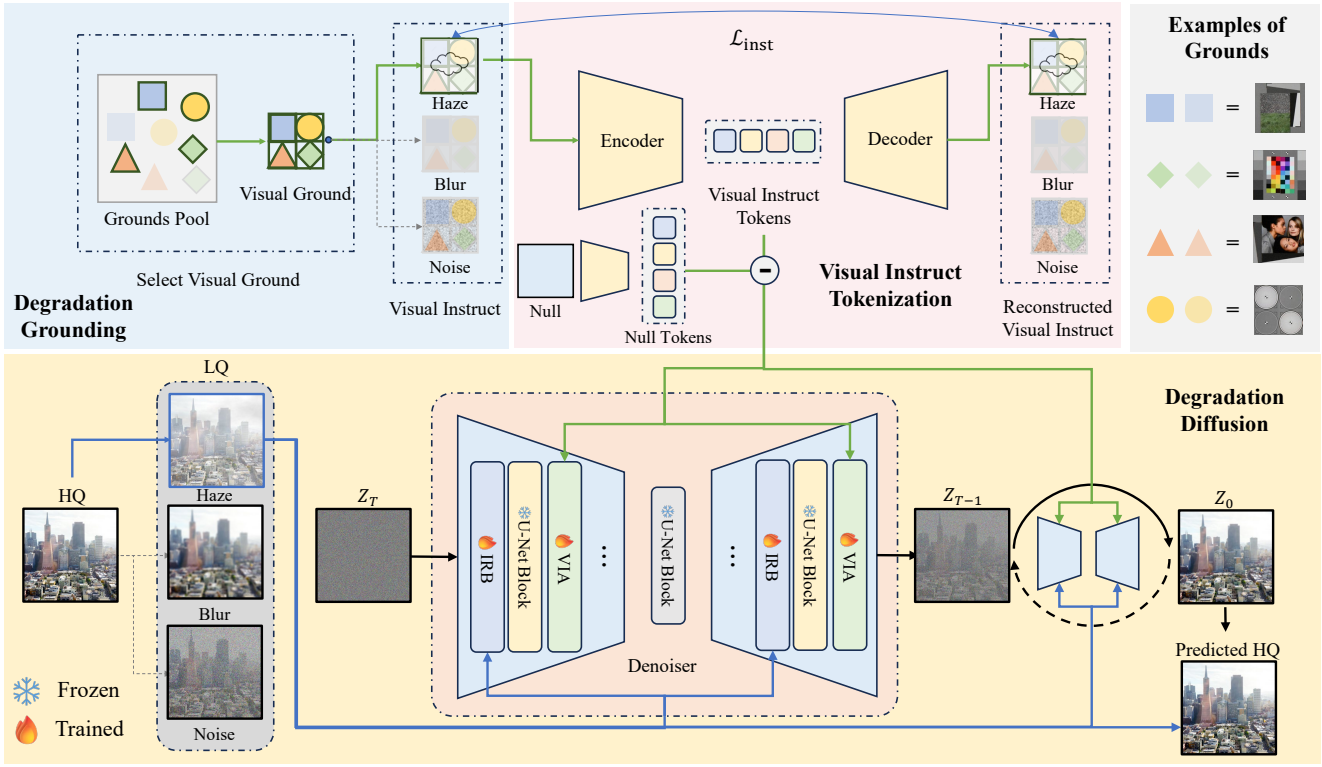


Figure 2. Framework of the proposed method. 1) Visual instructions are constructed from visual grounds to demonstrate the visual effects of the image degradations, 2) visual instructions are tokenized and contrasted with the "clean" (null) visual grounds, 3) finally, the visual instruction tokens guide the denoising diffusion model which estimates the degradation according to the hint of visual instructions. The visual instruction tokenizer is a quantized auto-encoder trained by the loss $\mathcal{L}_{\text{inst}}$, while only the encoder is used for inference.

to increased accumulated errors during inference. Moreover, they require training from scratch, unable to leverage prior knowledge about natural (clean) HQ images in pretrained models [94, 97, 101]. Instead, we diffuse the degradations between HQ and LQ conditioned on LQ images, guided by our visual instructions via the Visual Instruction Adapter (VIA). Our approach is simple yet effective for all-in-one image restoration, allowing for stable training and easy adaptation from pretrained models.

2.3. Visual Instructions in Diffusion Models

Previous works guide diffusion models to generate images with various types of instructions, including both texts [29, 56, 94, 97, 101] and visual instructions such as depth maps [140], segmentation maps [20, 117], layouts [97], inpainting masks [38, 127], sketches [116], skeletons [80], keypoints [55], example images [10, 82, 118], to name a few. Visual instructions are constructed to demonstrate image generation requirements or desired edits that are difficult to describe precisely in languages, such as turning the image into a drawing of a specific style [82]. In our work, we develop a novel visual instruction dedicated to image restoration, which more accurately captures the subtle low-level visual details of degradations than texts such as "noise" or "blurring", leading to better performance.

3. Method

This section introduces the proposed diffusion-based all-in-one image restoration method, Defusion. Fig. 2 depicts the overall framework. Sec. 3.1 describes the construction of our visual instructions. Sec. 3.2 formulates the instruction-guided diffusion model in degradation space. Sec. 3.3 details the conditioning mechanisms with the visual instructions via a Visual Instruction Adapter (VIA) and condition generation on the LQ images with an Image Restoration Bridge (IRB).

3.1. Visual Instructions

Motivation. Most all-in-one image restoration estimate degradation types given the input LQ images to guide restoration [4, 59, 84, 86], where the estimation becomes inaccurate for unseen degradation types. Inspired by InstructPix2Pix [12], some work [19, 43] control the restoration process with user-provided text instructions such as "make this blurry images sharper". However, *languages and visual patterns reside in very different spaces*. While the high-

level semantics can be aligned by visual-language pretraining [5, 61, 93], language cannot accurately convey visual details [79, 103] such as how the degradations rearrange and combine pixels.

On the contrary, we *let visual speak for visual* and resort to accurately describing the visual effect of degradations through visual instructions. Ideally, the visual instructions should capture the visually relevant aspects of degradations with minimal information loss and be disentangled with image semantics to improve generalizability. Next, we describe how to construct such visual instructions.

Grounding of Degradations. Image degradations are "dangling," meaning that their visual effects only manifest when they exist in the context of degraded images. Therefore, we first apply degradations on some "standard images" to visually demonstrate the degradations to the restoration model. We call this process *grounding* of degradations, the "standard images" are dubbed visual grounds.

The visual grounds are carefully chosen to include various types of visual constructs, patterns, structures, etc., that may occur in natural images to reveal the full extent of degradation. We draw inspiration from image quality assessment [77, 78, 114] and select TE42 [26], a family of charts commonly used for camera testing and visual analysis, to construct a pool of visual grounds. TE42 comprises a rich combination of textures and color palettes. We group the constructs in TE42 charts into concentric textures, random textures, natural textures, and colors and randomly select one construct from each group to form a visual ground. Fig. 2 shows some examples of visual grounds. Given an arbitrary degradation, we transform the visual ground by the degradation with random parameters. The degraded visual ground forms the visual instruction.

Degradation Tokenization. The grounded degradations (i.e. our visual instructions) obtained above can be regarded as "visual words" that "speak" how the degradations visually distort the image content. We further "tokenize" the grounded degradations into *degradation tokens* to guide diffusion-based image restoration, analogous to how text instructions are tokenized and guide text-to-image generation.

In our work, we tokenize our visual instructions with a Degradation Tokenizer that produces a sequence of discrete degradation token embeddings. Degradation Tokenizer compresses the perceptually unimportant constituents in the visual instructions with a quantization-regularized autoencoder inspired by LDM [97] and VQ-VAE [113].

Specifically, a visual instruction $\boldsymbol{v} \in \mathbb{R}^{H_v \times W_v \times 3}$ is mapped to embedding $\boldsymbol{e}_v \in \mathbb{R}^{h_v \times w_v \times C_v}$ (C_V is latent dimension) by an encoder $\boldsymbol{E}$. Embedding $\boldsymbol{e}_v$ is then vector-quantized (VQ) into one of the *degradation codes* $\boldsymbol{z}_v \in \mathbb{R}^{h_v \times w_v \times C_v}$ in the codebook $\mathcal{Z} = \{\boldsymbol{e}_k\}_{k=1}^K$ of size K by

$$\boldsymbol{z}_v = \arg\min_{\boldsymbol{e}_k \in \mathcal{Z}} \{\|\boldsymbol{e}_k - \boldsymbol{e}_v\|_2 = \|\boldsymbol{e}_k - \boldsymbol{E}(\boldsymbol{v})\|_2\}, \quad (1)$$

where $\|\cdot\|_2$ is Euclidean distance. The degradation code $\boldsymbol{z}_v$ is further reconstructed by a decoder $\boldsymbol{D}$ to approximate the original visual instruction $\hat{\boldsymbol{v}} = \boldsymbol{D}(\boldsymbol{z}_v)$.

Training. The encoder $\boldsymbol{E}$, decoder $\boldsymbol{D}$ and degradation codebook $\mathcal{Z}$ are jointly optimized by the following loss:

$$\mathcal{L}_{\text{inst}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{VQ}}, \quad (2)$$

$$\mathcal{L}_{\text{VQ}} = \|\text{sg}\left[\boldsymbol{E}(\boldsymbol{v}) - \boldsymbol{z}_v\right]\| + \|\boldsymbol{E}(\boldsymbol{v}) - \text{sg}\left[\boldsymbol{z}_v\right]\|, \quad (3)$$

$$\mathcal{L}_{\text{rec}} = \|\boldsymbol{D}(\boldsymbol{z}_v) - \boldsymbol{v}\|_2^2 + \lambda\, \mathcal{L}_{\text{LPIPS}}(\boldsymbol{D}(\boldsymbol{z}_v), \boldsymbol{v}), \quad (4)$$

where $\text{sg}[\cdot]$ is the stop-gradient operator [113], and the gradient back-propagates through $\boldsymbol{z}_v$ by straight-through estimator [11]. The reconstruction loss $\mathcal{L}_{\text{rec}}$ consists of a mean-squared error (MSE) and a perceptual loss $\mathcal{L}_{\text{LPIPS}}$ [142] to improve the perception of visual details, balanced by hyperparameter λ. To improve visual quality, we also adopt a hinge-based adversarial loss [27, 97]. We randomly replace the visual instruction $\boldsymbol{v}$ with its clean visual ground with a probability of 0.1 during training.

Inference. The trained encoder $\boldsymbol{E}$ serves as our Degradation Tokenizer. The visual instructions discretized and encoded by $\boldsymbol{E}$ capture the essential visual patterns and structures of degradations. To improve the discriminability of degradations, we also encode the clean visual grounds ("null" in Fig. 2) without degradation and subtract their embeddings from those of $\boldsymbol{v}$ and use the residuals to guide our degradation diffusion in Sec. 3.2.

3.2. Degradation Diffusion

Preliminary: Diffusion. Diffusion models [36, 46] generate data $\boldsymbol{x}_0 \sim p_0(\cdot)$ by reversing the following forward process that "diffuses" $\boldsymbol{x}_0$ into standard Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$:

$$p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}); \quad 0 \le t \le 1, \quad (5)$$

where α_t, σ_t define the noise schedule. The reversing is achieved by solving the reverse-time stochastic differential equation (SDE) [70, 108] of the reversed diffusion process:

$$\mathrm{d}\boldsymbol{x}_t = \left[f(t)\boldsymbol{x}_t - g^2(t)\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] \mathrm{d}t + g(t)\, \mathrm{d}\widetilde{\boldsymbol{w}}_t, \quad (6)$$

where $\widetilde{\boldsymbol{w}}_t$ is the reverse-time standard Wiener process, and

$$f(t) = \frac{\mathrm{d}\log \alpha_t}{\mathrm{d}t}, \quad g^2(t) = -\sigma_t^2 \frac{\mathrm{d}\log(\alpha_t/\sigma_t)}{\mathrm{d}t}. \quad (7)$$

A score model $\boldsymbol{s}_\theta(\boldsymbol{x_t}, t)$ is trained to estimate the score $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ in Eq. (6) via denoising score matching [115]:

$$\min_\theta \mathbb{E}_{t, \boldsymbol{\epsilon}, \boldsymbol{x}_0, \boldsymbol{x}_t} \left[\|\epsilon + \sigma_t \boldsymbol{s}_\theta(\boldsymbol{x_t}, t)\|_2^2\right]. \quad (8)$$

Degradation Diffusion. Previous works [4, 43, 59, 131] model the distribution of HQ images $\boldsymbol{x}_{\text{HQ}}$ with diffusion

models guided by LQ images $\boldsymbol{x}_{\text{LQ}}$, so that $\boldsymbol{x}_0 = \boldsymbol{x}_{\text{HQ}}$ in Eq. (5). We call them diffusion in image space. We found their formulation inefficient for instruction-guided image restoration, wherein our visual instructions primarily describe the degradations and are loosely correlated with HQ image semantics. Instead, we formulate our visual instruction guided diffusion model in *degradation space*.

Formally, denote $\mathcal{T}_{\boldsymbol{v}}(\boldsymbol{x}_{\text{LQ}})$ to be the "restored" result of LQ image $\boldsymbol{x}_{\text{LQ}}$ induced by visual instruction $\boldsymbol{v}$. That is, $\mathcal{T}_{\boldsymbol{v}}(\boldsymbol{x}_{\text{LQ}})$ is the "cleaner" version of $\boldsymbol{x}_{\text{LQ}}$ where the degradations are removed *if and only if* they are indicated by $\boldsymbol{v}$, and $\mathcal{T}_{\boldsymbol{v}}(\boldsymbol{x}_{\text{LQ}}) = \boldsymbol{x}_{\text{HQ}}$ when $\boldsymbol{v}$ describes all degradations in $\boldsymbol{x}_{\text{LQ}}$. The degradation space $\mathcal{D}_{\boldsymbol{v}}$ induced by $\boldsymbol{v}$ is defined as the set of difference vectors between LQ image set $\mathcal{S}_{\text{LQ}}$ and their corresponding $\mathcal{T}_{\boldsymbol{v}}(\boldsymbol{x}_{\text{LQ}})$:

$$\mathcal{D}_{\boldsymbol{v}} := \left\{ \boldsymbol{x}_{\text{LQ}} - \mathcal{T}_{\boldsymbol{v}}(\boldsymbol{x}_{\text{LQ}}) \,\middle|\, \boldsymbol{x}_{\text{LQ}} \in \mathcal{S}_{\text{LQ}} \right\}. \tag{9}$$

The degradation diffusion diffuses $\mathcal{D}_{\boldsymbol{v}}$ into the support set of standard Gaussian noise $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The degradation diffusion is conditioned on both LQ images (HQ images are unavailable during inference) and visual instruction $\boldsymbol{v}$:

$$\begin{aligned} &p_t(\boldsymbol{y}_t|\boldsymbol{y}_0, \boldsymbol{x}_{\text{LQ}}, \boldsymbol{v}) = \mathcal{N}(\alpha_t \boldsymbol{y}_0, \sigma_t^2 \boldsymbol{I}) \\ &\text{where } \boldsymbol{y}_0 = \boldsymbol{x}_{\text{LQ}} - \mathcal{T}_{\boldsymbol{v}}(\boldsymbol{x}_{\text{LQ}}),\ 0 \le t \le 1. \end{aligned} \tag{10}$$

We train the conditioned score model $\boldsymbol{s}_\theta(\boldsymbol{y_t}, t, \boldsymbol{x}_{\text{LQ}}, \boldsymbol{v})$ to estimate the conditioned score $\nabla_{\boldsymbol{y}_t} \log p_t(\boldsymbol{y}_t|\boldsymbol{x}_{\text{LQ}}, \boldsymbol{v})$ via the following objective:

$$\min_\theta \mathbb{E}_{t, \boldsymbol{\epsilon}, \boldsymbol{x}_{\text{LQ}}, \boldsymbol{v}, \boldsymbol{x}_t} \left[\|\epsilon + \sigma_t \boldsymbol{s}_\theta(\boldsymbol{y_t}, t, \boldsymbol{x}_{\text{LQ}}, \boldsymbol{v})\|_2^2 \right]. \tag{11}$$

After training, we numerically solve the reverse time SDE for $t = 1 \to 0$ from $\boldsymbol{y}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to obtain $\hat{\boldsymbol{y}}_0$ as

$$\mathrm{d}\boldsymbol{y}_t = \left[f(t)\boldsymbol{y}_t - g^2(t)\boldsymbol{s}_\theta(\boldsymbol{y_t}, t, \boldsymbol{x}_{\text{LQ}}, \boldsymbol{v}) \right] \mathrm{d}t + g(t)\, \mathrm{d}\widetilde{\boldsymbol{w}}_t, \tag{12}$$

where f, g are defined by Eq. (7). The final restored image is then computed as $\hat{\mathcal{T}}_{\boldsymbol{v}}(\boldsymbol{x}_{\text{LQ}}) = \boldsymbol{x}_{\text{LQ}} - \hat{\boldsymbol{y}}_0$.

Our diffusion formulation in degradation space has several advantages over previous diffusion in image space [4, 43, 59, 131]: 1) it achieves better restoration because $\boldsymbol{y}_t$ is highly correlated with degradations; 2) training is more stable since the distribution of $\boldsymbol{y}_t$ is more consistent than images', which also reduces required model capacity; 3) our formulation *does nothing* when $\boldsymbol{v}$ does not match the actual degradation in LQ images, leading to better controllability. Besides, unlike [63, 105, 132], our formulation has no special requirement for α_t, σ_t; thus, we can adapt any pretrained diffusion model into our formulation with minimal efforts.

3.3. Conditioning

The score model $\boldsymbol{s}_\theta(\boldsymbol{y_t}, t, \boldsymbol{x}_{\text{LQ}}, \boldsymbol{v})$ trained by Eq. (11) is conditioned on the LQ image $\boldsymbol{x}_{\text{LQ}}$ and visual instruction $\boldsymbol{v}$. We implement $\boldsymbol{s}_\theta$ with U-Net [98] and introduce our conditioning mechanisms in this section.

Image Restoration Bridge (IRB). The IRB introduces the conditioning on LQ images. IRB first embeds LQ images into hierarchical feature maps of different resolutions with a lightweight convolutional network [80]; each resolution level of feature maps corresponds to that of the diffusion U-Net. The features are injected into the U-Net via AdaLN-Zero [85]. We found no benefit in concatenating the LQ image with the U-Net input [12]. Unlike previous works [4, 59, 131], the performance deteriorates if we inject the features into the U-Net decoder with ControlNet [140] or a similar adapter. We conjecture the reason is that the prediction in degradation space is dissimilar to the LQ images, leading to diverging features in the U-Net encoder.

Visual Instruction Adapter (VIA). Visual instructions control the diffusion U-Net through the VIA. Inspired by IP-Adapter [129], we copy the cross-attention layer in the pretrained text-to-image diffusion U-Net. The query of the copied cross-attention is the (flattened) U-Net feature map, and the key and value are the degradation codes $\boldsymbol{z}_v$ encoded from the visual instructions $\boldsymbol{v}$ (see Sec. 3.1). This preserves the text-conditioning ability of the pretrained model, allowing the injection of text descriptions as additional guidance.

4. Experiments

4.1. Experiment Settings

Datasets and Metrics. We combined various image restoration datasets to form the training set, following [137]. This includes datasets for deraining such as Rain1400 [30], Outdoor-Rain [54], and LHP [33]; dehazing datasets like RESIDE [49], and Dense-Haze [6]; desnowing datasets such as Snow100K [67] and RealSnow [149]; raindrop removal datasets like RainDrop [87] and RainDS [90]; real denoising with SIDD [1]; motion deblurring with GoPro [81] and RealBlur [96]; defocus deblurring with DPDD [2].

During the evaluation, we used Dense-Haze [6] for dehazing, Rain1400 [30] for deraining, GoPro [81] for motion deblurring, RainDrop [87] for raindrop removal, DPDD [2] for defocus deblurring, Snow100K [67] for desnowing, SIDD [1] for real denoising, and LIVE1 [104] for JPEG artifact removal (with quality factor $q = 10$) on the synthetic datasets. For real-world degradation datasets, we evaluated our model on the NH-HAZE [7] dataset, LHP [33] dataset, RealBlur [96] dataset, and the RealSnow [149] dataset.

Moreover, following the method in [84], we constructed a mixed distortion dataset on the WED [75] dataset using rain, snow, and noise distortions. We used these diverse relationships in this dataset to evaluate the model's ability to handle complex distortions. We also introduced underwater datasets EUVP [39] and TURBID [24] to examine the

Table 1. Comparison with state-of-the-art task-specific methods and all-in-one methods on 8 tasks. General IR models trained under the all-in-one setting are marked with a symbol *. The best and second-best performances are in red and bold font, with the top 3 with a light black background.

Motion Deblur (Gopro [81])			**Defocus Deblur (DPDD [2])**			**Desnowing (Snow100K-L [67])**			**Dehazing (Dense-Haze [6])**		
Method	PSNR	SSIM	Method	PSNR	SSIM	Method	PSNR	SSIM	Method	PSNR	SSIM
Task Specific											
MPRNet[133]	32.66	0.959	DRBNet[99]	25.73	0.791	DesnowNet[67]	27.17	0.898	AECRNet[123]	15.80	0.466
Restormer[134]	32.92	0.961	Restormer[134]	25.98	0.811	DDMSNet[138]	28.85	0.877	FocalNet[21]	**17.07**	**0.630**
Stripformer[111]	33.08	0.962	NRKNet[92]	26.11	0.810	DRT[58]	29.56	0.892	DeHamer[32]	16.62	0.560
DiffIR[125]	**33.20**	**0.963**	FocalNet[21]	26.18	0.808	WeatherDiff[83]	30.43	0.915	MB-Taylor[89]	16.44	0.566
All in One											
AirNet[50]	28.31	0.910	AirNet[50]	25.37	0.770	AirNet[50]	30.14	0.907	AirNet[50]	15.32	0.562
PromptIR[86]	31.02	0.938	PromptIR[86]	25.66	0.791	PromptIR[86]	30.91	0.913	PromptIR[86]	16.71	0.580
DA-CLIP[71]	27.12	0.823	DA-CLIP[71]	24.91	0.749	DA-CLIP[71]	28.31	0.862	TransWeather[112]	15.58	0.569
Restormer*[134]	31.17	0.931	Restormer*[134]	28.08	**0.898**	Restormer*[134]	29.18	0.899	Restormer*[134]	16.18	0.560
NAFNet*[15]	31.44	0.939	NAFNet*[15]	**28.21**	0.892	NAFNet*[15]	29.21	0.896	NAFNet*[15]	16.72	0.602
MPerceiver[4]	32.49	0.959	MPerceiver[4]	25.88	0.803	MPerceiver[4]	**31.02**	**0.916**	WGWS-Net[139]	15.50	0.535
Defusion(Ours)	**34.53**	**0.966**	**Defusion(Ours)**	**29.68**	**0.922**	**Defusion(Ours)**	**32.11**	**0.926**	**Defusion(Ours)**	**17.55**	**0.677**
Raindrop Removal (RainDrop [87])			**Deraining (Rain1400 [30])**			**Real Denoising (SIDD [1])**			**Artifact removal (LIVE1 [104])**		
Method	PSNR	SSIM	Method	PSNR	SSIM	Method	PSNR	SSIM	Method	PSNR	SSIM
Task Specific											
AttentGAN[87]	31.59	0.917	Uformer[120]	32.84	0.931	MPRNet[133]	39.71	0.958	QGAC[25]	27.62	0.804
Quanetal.[91]	31.37	0.918	Restormer[134]	**33.68**	0.939	Uformer[120]	39.89	0.960	FBCNN[18]	27.77	0.803
IDT[126]	31.87	0.931	DRSformer[17]	33.66	0.939	Restormer[134]	40.02	0.960	SwinIR[128]	27.45	0.796
UDR-S[16]	32.64	0.942	UDR-S2[16]	33.08	0.930	ART[136]	39.99	0.960	DAGN[76]	**27.95**	0.807
All in One											
AirNet[50]	31.32	0.925	AirNet[50]	32.36	0.928	AirNet[50]	38.32	0.945	AirNet[50]	15.32	0.562
PromptIR[86]	32.03	0.938	PromptIR[86]	33.26	0.935	PromptIR[86]	39.52	0.954	TransWeather[112]	15.58	0.569
DA-CLIP[71]	30.44	0.880	DA-CLIP[71]	29.67	0.851	DA-CLIP[71]	34.04	0.824	WGWS-Net[139]	15.50	0.535
Restormer*[134]	31.67	**0.958**	Restormer*[134]	32.61	**0.944**	Restormer*[134]	**40.60**	0.955	Restormer*[134]	26.77	**0.827**
NAFNet*[15]	31.03	0.952	NAFNet*[15]	32.38	0.943	NAFNet*[15]	40.22	0.952	NAFNet*[15]	26.79	**0.827**
MPerceiver[4]	**33.21**	0.929	MPerceiver[4]	33.40	0.937	MPerceiver[4]	39.96	0.959	MPerceiver[4]	**27.79**	0.804
Defusion(Ours)	**33.81**	**0.967**	**Defusion(Ours)**	**34.09**	**0.954**	**Defusion(Ours)**	**40.89**	0.954	**Defusion(Ours)**	27.77	**0.832**

ability of our method to deal with composite distortions in special environments. We used PSNR and SSIM [119] as distortion metrics, LPIPS [142] as perceptual metrics.

Implementation Details. Defusion was trained using the Adam optimizer [48] with learning rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, starting with an initial learning rate of 1e-4 and decaying to 1e-6 via cosine annealing [69]. We trained for 100 epochs on the mixed dataset using 8 NVIDIA A100 GPUs at a resolution of 256x256. Due to varying data volumes for different tasks, we adjusted the frequency of random patch cropping across datasets to balance data quantity. For data augmentation, we applied horizontal and vertical flips to all datasets. The model used the U-Net [22] architecture, DDIM [107] as the sampling strategy during inference, and 4 timesteps for all tasks.

4.2. Comparison with state-of-the-art methods

To comprehensively evaluate the effectiveness of Defusion, we compared it with general architecture-based image restoration methods and unified frameworks. Tab. 1 compares performance with state-of-the-art (SOTA) methods across 8 tasks. Across all datasets, Defusion outperformed other unified methods. As a unified model, Defusion even surpasses task-specific methods in several cases. Compared to text-based instructions [71], ours achieved significant improvements across all tasks. Our method also outperforms learned task embeddings [4, 86], sometimes by large margins. Defusion also achieves best perceptual qualities (see the supplementary material).

Fig. 3 and Fig. 4 show qualitative results on various datasets. In Fig. 3, our method demonstrates superior de-

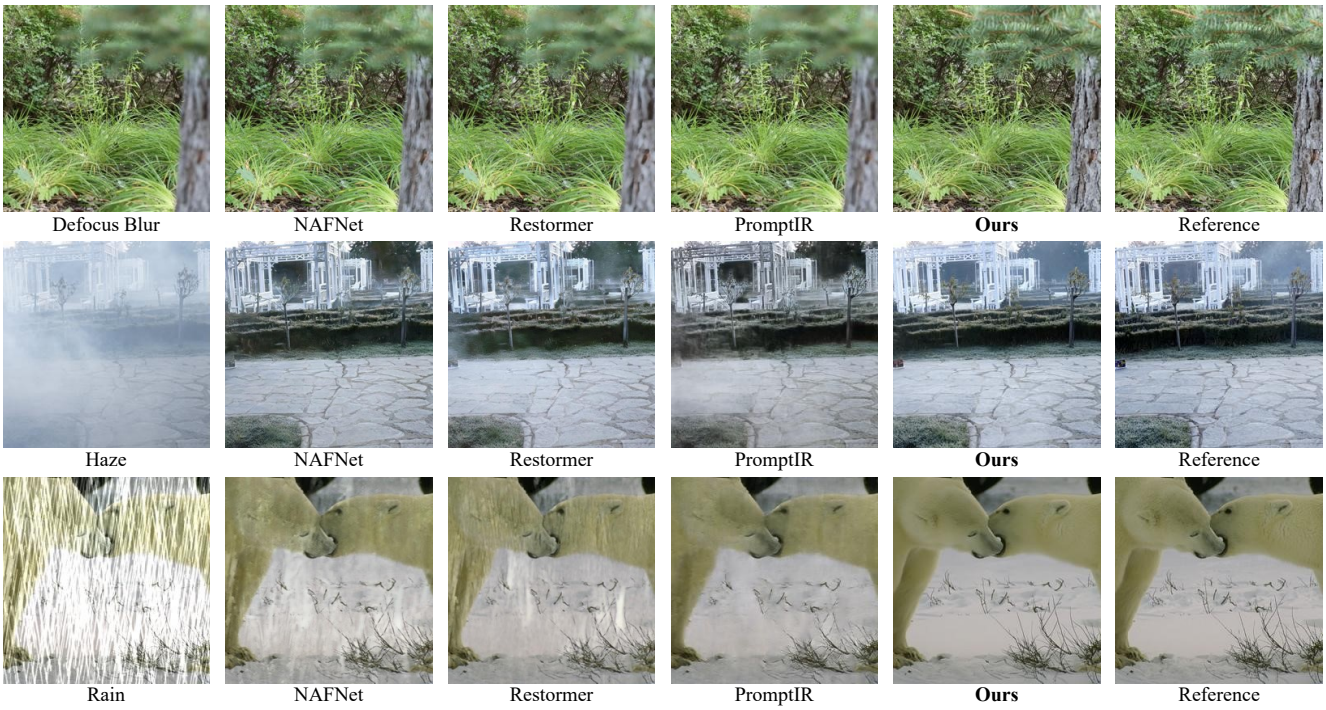


Figure 3. Visual results on dehazing, draining, and deblurring on synthetic datasets .

Table 2. Real-world datasets results. General IR models trained under the all-in-one training set are marked with a symbol *. The best and second-best performances are in red and bold font, with the top 3 with a light black background.

Dehazing (NH-HAZE [7])			**Deraining (LHP [33])**			**Desnowing (RealSnow [149])**			**Motion Deblur (RealBlur-J [96])**		
Method	PSNR	SSIM	Method	PSNR	SSIM	Method	PSNR	SSIM	Method	PSNR	SSIM
Task Specific											
AECRNet[123]	19.88	0.625	MUSS[37]	30.02	0.886	MIRNetv2[135]	31.39	0.916	MPRNet[133]	28.70	0.873
SGID[9]	16.23	0.606	Restormer[134]	29.72	0.889	ART[136]	31.05	0.913	Restormer[134]	28.96	0.879
DeHamer[32]	20.66	0.684	DRSformer[17]	30.04	0.895	Restormer[134]	31.38	0.923	Stripformer[111]	28.82	0.876
MB-Taylor[89]	19.43	0.638	UDR-S2[16]	28.59	0.884	NAFNet[15]	31.44	0.919	DiffIR[125]	29.06	0.882
All in One											
AirNet[50]	17.63	0.615	AirNet[50]	31.73	0.889	AirNet[50]	31.02	0.923	AirNet[50]	27.91	0.834
TransWeather[112]	19.19	0.644	TransWeather[112]	29.87	0.867	TransWeather[112]	31.13	0.922	TransWeather[112]	28.03	0.837
WGWS-Net[139]	18.39	0.639	WGWS-Net[139]	30.77	0.885	WGWS-Net[139]	31.37	0.919	WGWS-Net[139]	28.10	0.838
Restormer*[134]	**22.57**	**0.795**	Restormer*[134]	29.87	**0.928**	Restormer*[134]	29.47	0.889	Restormer*[134]	**30.91**	**0.930**
NAFNet*[15]	21.66	0.783	NAFNet*[15]	27.95	0.917	NAFNet*[15]	28.15	0.858	NAFNet*[15]	30.80	0.928
MPerceiver[4]	21.97	0.646	MPerceiver[4]	**32.07**	0.889	MPerceiver[4]	**31.45**	**0.924**	MPerceiver[4]	29.13	0.881
Defusion(Ours)	**22.87**	**0.853**	**Defusion(Ours)**	**32.98**	**0.947**	**Defusion(Ours)**	**32.73**	**0.930**	**Defusion(Ours)**	**32.39**	**0.945**

hazing. In the deraining task (Fig. 3), our method more effectively removes rain streaks and restores textures in occluded regions. We achieved sharper results with fewer artifacts for deblurring (Fig. 3). In the real-world datasets, We provide better de-blurring and de-snowing performance (Fig. 4). For simplicity, visual results for other tasks are provided in the supplementary material. Additionally, Tab. 3 compares the mixed distortion dataset involving rain, snow, and noise. Our method demonstrated a significant advantage in handling challenging mixed-distortion images. Tab. 4 provides results in an underwater dataset, which exemplify the superior performance of our method in the face of natural mixing distortions.

4.3. Ablation Study

In this section, we conduct ablation studies on visual grounding, visual instruction embedding, and degradation diffusion to demonstrate the effectiveness of the different components proposed in Defusion. Naive diffusion is used to exclude diffusion performance in degradation space.



Figure 4. Real-world visual results on desnowing and deblurring.

Table 3. Synthesized mixed distortion dataset. The → indicates the order in which distortions are added. General IR models trained under the all-in-one training set are marked with a symbol *.

Method	**Rain**→Noise→Snow		**Rain**→Snow→ Noise		**Noise**→Snow→ Rain		**Noise**→Rain→ Snow		**Snow**→Rain→ Noise		**Snow**→Noise→ Rain	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
WeatherDiff[83]	14.744	0.363	14.693	0.361	14.913	0.397	14.839	0.393	14.688	0.359	14.843	0.395
PromptIR[86]	13.976	0.471	14.212	0.492	14.544	0.504	14.357	0.494	14.553	0.510	14.615	0.513
Restormer*[134]	24.114	0.766	23.895	0.760	25.168	0.802	25.022	0.798	24.144	0.766	24.814	0.793
NAFNet*[15]	23.815	0.755	23.775	0.755	24.203	0.776	24.204	0.773	23.564	0.745	24.106	0.772
Defusion(Ours)	**25.511**	**0.812**	**25.228**	**0.807**	**27.008**	**0.855**	**26.600**	**0.847**	**25.052**	**0.802**	**26.079**	**0.837**

Table 4. UnderWater datasets results. Method trained under the all-in-one training set are marked with a symbol *.

Method	**EUVP** [39]			**TURBID** [24]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
PromptIR[86]	20.046	0.799	0.296	15.324	0.593	0.324
Restormer*[134]	19.764	0.788	0.316	17.420	0.648	0.282
NAFNet*[15]	19.661	0.779	0.314	17.571	0.672	0.291
Defusion(Ours)	**21.233**	**0.814**	**0.269**	**18.261**	**0.659**	**0.283**

Table 5. Ablation study. The metrics are reported on the average of deraining, dehazing and desnowing.

Method	**Average**		
	PSNR↑	SSIM↑	LPIPS↓
Defusion(Ours)	33.83	0.951	0.025
- replace with Naive Diffusion	31.72	0.918	0.086
- replace with Blank Grounding	30.89	0.903	0.102
- replace with Simple Grounding	31.64	0.927	0.044
- replace with Text Instruction	32.57	0.929	0.041
- without Instruction	30.86	0.903	0.101

Blank grounding consists of a blank image, and simple grounding is a simple striped image. Both are used to test the high response of Grounding to distortion. Text and non-instruction are used to compare the performance of Visual Instruction instructions. The results demonstrate the excellent performance of our visual instruction and degradation diffusion designs.

5. Conclusion

We present **Defusion**, a novel image restoration framework utilizing visual instruction-guided degradation diffusion. Defusion aligns visual instructions with degradation patterns, surpassing traditional methods dependent on ambiguous text-based priors or task-specific models. Operating in degradation space, our model reconstructs high-quality images by denoising the shift between LQ and HQ images, enhancing stability and performance. Extensive experiments demonstrate Defusion's state-of-the-art performance in various image restoration tasks, including complex real-world scenarios with multiple degradations. The proposed approach sets new state-of-the-art for multitask image restoration with a unified model based on generative models, providing a robust solution to handle a wide range of restoration applications in the real world.

Acknowledgments This work was supported by Beijing Natural Science Foundation (JQ24022), CAAI-Ant Group Research Fund CAAI-MYJJ 2024-02, the National Natural Science Foundation of China (No. 62372451, No. 62192785, No. 62372082, No. 62403462), the Key Research and Development Program of Xinjiang Uyghur Autonomous Region, Grant No. 2023B03024.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018. 5, 6, 2, 3

[2] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 111–126. Springer, 2020. 5, 6, 1, 3

[3] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 2

[4] Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25432–25444, 2024. 1, 2, 3, 4, 5, 6, 7

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022. 4

[6] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *2019 IEEE international conference on image processing (ICIP)*, pages 1014–1018. IEEE, 2019. 5, 6, 2

[7] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 444–445, 2020. 5, 7

[8] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3155–3164, 2019. 1

[9] Haoran Bai, Jinshan Pan, Xinguang Xiang, and Jinhui Tang. Self-guided image dehazing using progressive feature fusion. *IEEE Transactions on Image Processing*, 31:1217–1229, 2022. 7

[10] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 3

[11] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 4

[12] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3, 5

[13] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26540–26550, 2024. 1

[14] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 1

[15] 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. 2, 6, 7, 8

[16] 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. 1, 6, 7, 3

[17] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023. 1, 6, 7, 3

[18] 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. 6

[19] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*, 2(4), 2024. 1, 2, 3

[20] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[21] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13001–13011, 2023. 6, 3

[22] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 6, 4

[23] Zheng Ding, Xuaner Zhang, Zhuowen Tu, and Zhihao Xia. Restoration by generation with constrained priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2567–2577, 2024. 2

[24] Amanda Duarte, Felipe Codevilla, Joel De O Gaya, and Silvia SC Botelho. A dataset to evaluate underwater image restoration methods. In *OCEANS 2016-Shanghai*, pages 1–6. IEEE, 2016. 5, 8, 3

[25] Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. Quantization guided jpeg artifact correction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 293–309. Springer, 2020. 1, 6

[26] Image Engineering. TE42. https://www.image-engineering.de/products/charts/all/425-te42, 2024. [Online; accessed 19-October-2024]. 4, 1

[27] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4, 3

[28] Yingying Fang, Hao Zhang, Hok Shing Wong, and Tieyong Zeng. A robust non-blind deblurring method using deep denoiser prior. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 734–743, 2022. 2

[29] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[30] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017. 5, 6, 2, 3

[31] Chunle Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5810, 2022. 2

[32] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5812–5820, 2022. 6, 7

[33] Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12097–12107, 2023. 5, 7, 2

[34] Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. *arXiv preprint arXiv:2407.04621*, 2024. 2

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. 4

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[37] Huaibo Huang, Mandi Luo, and Ran He. Memory uncertainty learning for real-world single image deraining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3446–3460, 2022. 7

[38] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 3

[39] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020. 5, 8, 3

[40] 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. 4

[41] Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2021. 1

[42] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020. 1

[43] Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. *arXiv preprint arXiv:2310.10123*, 2023. 1, 2, 3, 4, 5

[44] Xin Jin, Linghao Han, Z. Li, Chunle Guo, Zhi Chai, and Chongyi Li. Dnf: Decouple and feedback network for seeing in the dark. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18135–18144, 2023. 2

[45] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 2

[46] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 4

[47] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[48] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6, 4

[49] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 5, 2

[50] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17452–17462, 2022. 6, 7, 3

[51] Bingchen Li, Xin Li, Yiting Lu, Ruoyu Feng, Mengxi Guo, Shijie Zhao, Li Zhang, and Zhibo Chen. Promptcir: Blind compressed image restoration with prompt learning. *arXiv preprint arXiv:2404.17433*, 2024. 2

[52] Chongyi Li, Chunle Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. *ArXiv*, abs/2302.11831, 2023. 2

[53] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *European conference on computer vision*, pages 736–753. Springer, 2022. 2

[54] 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, 2

[55] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[56] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. 3

[57] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2

[58] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 589–598, 2022. 6, 3

[59] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 1, 2, 3, 4, 5

[60] Xin Yu Lin, Chao Ren, Xiao Liu, Jie Huang, and Yinjie Lei. Unsupervised image denoising in real-world scenarios via self-collaboration parallel generative adversarial branches. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12608–12618, 2023. 2

[61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[62] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8038–8047, 2024. 2

[63] Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2773–2783, 2024. 2, 5

[64] 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

[65] Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, and Rynson WH Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4208, 2024. 1, 2

[66] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 1

[67] 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, 6, 1, 3

[68] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[69] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6, 4

[70] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787, 2022. 4

[71] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 3(8), 2023. 6, 3

[72] 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

[73] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691, 2023. 2

[74] Jiaqi Ma, Tianheng Cheng, Guoli Wang, Qian Zhang, Xinggang Wang, and Lefei Zhang. Prores: Exploring degradation-aware visual prompt for universal image restoration. *arXiv preprint arXiv:2306.13653*, 2023. 2

[75] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016. 5, 2

[76] Li Ma, Yifan Zhao, Peixi Peng, and Yonghong Tian. Sensitivity decouple learning for image compression artifacts reduction. *IEEE Transactions on Image Processing*, 2024. 6

[77] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial

domain. *IEEE Transactions on image processing*, 21(12): 4695–4708, 2012. 4, 1

[78] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 4, 1

[79] Amin Karimi Monsefi, Kishore Prakash Sailaja, Ali Alilooee, Ser-Nam Lim, and Rajiv Ramnath. Detailclip: Detail-oriented clip for fine-grained tasks. *arXiv preprint arXiv:2409.06809*, 2024. 4

[80] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3, 5

[81] 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. 5, 6, 1, 3

[82] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[83] 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*, 45(8):10346–10357, 2023. 6, 8, 3

[84] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5815–5824. IEEE, 2023. 1, 2, 3, 5

[85] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5

[86] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 6, 8

[87] 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. 5, 6, 2, 3

[88] Chu-Jie Qin, Rui-Qi Wu, Zikun Liu, Xin Lin, Chun-Le Guo, Hyun Hee Park, and Chongyi Li. Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration. *arXiv preprint arXiv:2409.19403*, 2024. 2

[89] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12802–12813, 2023. 6, 7

[90] 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, 2

[91] 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. 6, 3

[92] Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2023. 6, 3

[93] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[94] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[95] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8596–8606, 2021. 1

[96] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 5, 7, 1

[97] 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. 2, 3, 4

[98] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 5

[99] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16304–16313, 2022. 6, 3

[100] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[101] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[102] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4

[103] Jie-Jing Shao, Jiang-Xin Shi, Xiao-Wen Yang, Lan-Zhe Guo, and Yu-Feng Li. Investigating the limitation of clip models: The worst-performing categories. *arXiv preprint arXiv:2310.03324*, 2023. 4

[104] H Sheikh. Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 5, 6, 2

[105] Zhenning Shi, Haoshuai Zheng, Chen Xu, Changsheng Dong, Bin Pan, Xueshuo Xie, Along He, Tao Li, and Huazhu Fu. Resfusion: Denoising diffusion probabilistic models for image restoration based on prior residual noise. *arXiv e-prints*, pages arXiv–2311, 2023. 2, 5

[106] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[107] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[108] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 4

[109] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 1

[110] 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. 1

[111] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1, 6, 7, 3

[112] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022. 6, 7

[113] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 4

[114] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE, 2015. 4, 1

[115] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 4

[116] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[117] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 3

[118] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 3

[119] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6, 4

[120] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 6, 3

[121] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2027–2036, 2022. 1

[122] 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. 1

[123] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10551–10560, 2021. 6, 7

[124] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 1

[125] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023. 2, 6, 7, 3

[126] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12978–12995, 2022. 6, 3

[127] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 3

[128] Wenhan Yang, Jiaying Liu, Shuai Yang, and Zongming Guo. Scale-free single image deraining via visibility-enhanced recurrent wavelet learning. *IEEE Transactions on Image Processing*, 28(6):2948–2961, 2019. 6

[129] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 5

[130] Tian Ye, Sixiang Chen, Wenhao Chai, Zhaohu Xing, Jing Qin, Ge Lin, and Lei Zhu. Learning diffusion texture priors for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2024. 2

[131] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 1, 2, 4, 5

[132] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[133] 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. 6, 7, 3

[134] 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. 2, 6, 7, 8, 3

[135] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1934–1948, 2022. 7

[136] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022. 6, 7, 3

[137] Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5825–5835, 2023. 5

[138] 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. 6, 3

[139] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021. 6, 7

[140] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 5

[141] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[142] 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. 4, 6

[143] Xu Zhang, Jiaqi Ma, Guoli Wang, Qian Zhang, Huan Zhang, and Lefei Zhang. Perceive-ir: Learning to perceive degradation better for all-in-one image restoration. *arXiv preprint arXiv:2408.15994*, 2024. 1, 2

[144] Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25827–25836, 2024. 2

[145] Zhaoyang Zhang, Yitong Jiang, Wenqi Shao, Xiaogang Wang, Ping Luo, Kaimo Lin, and Jinwei Gu. Real-time controllable denoising for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14028–14038, 2023. 2

[146] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8281–8291, 2024. 2

[147] Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25445–25455, 2024. 2

[148] Naishan Zheng, Man Zhou, Yanmeng Dong, Xiangyu Rui, Jie Huang, Chongyi Li, and Feng Zhao. Empowering low-light image enhancer through customized learnable priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12559–12569, 2023. 2

[149] Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21747–21758, 2023. 5, 7, 1