Multimodal Prompt Perceiver: Empower Adaptiveness, Generalizability and Fidelity for All-in-One Image Restoration

Yuang Ai[1,2] Huaibo Huang[1,2✉] Xiaoqiang Zhou[1,3] Jiexiang Wang[1,3] Ran He[1,2]
[1]MAIS & CRIPAC, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]University of Science and Technology of China, Hefei, China
shallowdream555@gmail.com, huaibo.huang@cripac.ia.ac.cn,
{xq525,jiexiang}@mail.ustc.edu.cn, rhe@nlpr.ia.ac.cn

Figure 1. Our **MPerceiver** excels in image restoration tasks with: **(I) All-in-one:** Addressing diverse degradations, including challenging mixed ones, through a single pretrained network. **(II) Zero-shot:** Handling training-unseen degradations effortlessly. **(III) Few-shot:** Adapting to new tasks with minimal data (about 3%-5% of data used by task-specific methods).

Abstract

Despite substantial progress, all-in-one image restoration (IR) grapples with persistent challenges in handling intricate real-world degradations. This paper introduces ***MPerceiver****: a novel multimodal prompt learning approach that harnesses Stable Diffusion (SD) priors to enhance adaptiveness, generalizability and fidelity for all-in-one image restoration. Specifically, we develop a dual-branch module to master two types of SD prompts: textual for holistic representation and visual for multiscale detail representation. Both prompts are dynamically adjusted by degradation predictions from the CLIP image encoder, enabling adaptive responses to diverse unknown degradations. Moreover, a plug-in detail refinement module improves restoration fidelity via direct encoder-to-decoder information transformation. To assess our method, MPerceiver is trained on 9 tasks for all-in-one IR and outper-*

✉Corresponding author. Project Page.

forms state-of-the-art task-specific methods across many tasks. Post multitask pre-training, MPerceiver attains a generalized representation in low-level vision, exhibiting remarkable zero-shot and few-shot capabilities in unseen tasks. Extensive experiments on 16 IR tasks underscore the superiority of MPerceiver in terms of adaptiveness, generalizability and fidelity.

1. Introduction

Image restoration (IR) aims to reconstruct a high-quality (HQ) image from its degraded low-quality (LQ) counterpart. Recent deep learning-based IR approaches excel in addressing single degradation, such as denoising [111, 127, 128, 131], deblurring [88, 94, 103], adverse weather removal [14, 36, 37, 47, 68, 81, 116], low-light enhancement [28, 114, 118], *etc*. However, these task-specific methods often fall short in real-world scenarios, such as autonomous driving and outdoor surveillance, where images may encounter unknown, dynamic degradations [73, 144]. The concept of all-in-one image restoration has recently gained significant traction, aiming to tackle multiple degradations with a unified model using a single set of pre-trained weights. Leading approaches leverage techniques such as contrastive learning [13, 49], task-specific sub-networks [82, 142], task-specific priors [104, 116], and task-agnostic priors [63] to enhance the network's capability across various degradations. Despite their promising performance, the adaptability and generalizability of all-in-one models to real-world scenarios, characterized by intricate and diverse degradations, remain challenging.

Large-scale text-to-image diffusion models, like Stable Diffusion (SD) [93], succeed in high-quality and diverse image synthesis. This motivates our exploration of leveraging SD for all-in-one image restoration, capitalizing on its HQ image priors to enhance reconstruction quality and generalization across realistic scenarios. However, direct application of SD faces challenges in adaptiveness, generalizability, and fidelity for all-in-one image restoration. SD's proficiency in HQ image synthesis relies on intricately designed prompts, complicating the crafting of suitable prompts for complex, authentic degradations, thereby limiting adaptability and generalization. Furthermore, as a latent diffusion model, SD adopts a VAE architecture with high compression, risking the loss of fine details in restored images and consequently restricting the fidelity of IR [18, 145].

In this paper, we propose MPerceiver, a multimodal prompt learning approach harnessing the generative priors of Stable Diffusion to enhance adaptiveness, generalizability and fidelity of all-in-one image restoration. MPerceiver comprises two modules: a dual-branch module learning textual and visual prompts for diverse degradations, and a detail refinement module (DRM) to boost restoration fidelity.

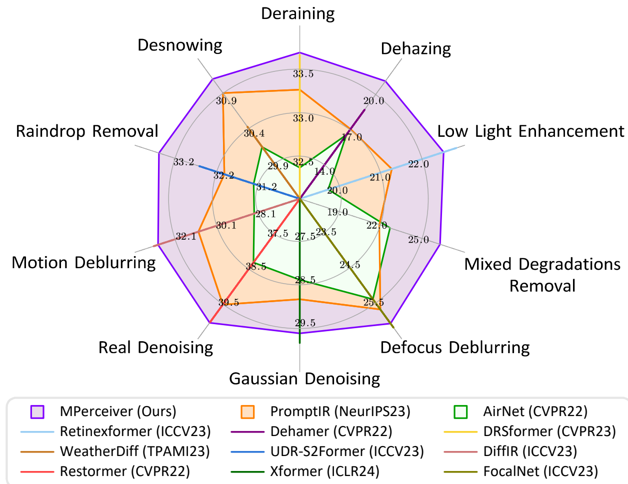


Figure 2. PSNR comparison with state-of-the-art all-in-one and task-specific methods across 10 tasks. Best viewed in color.

For textual prompt learning, it predicts HQ text embeddings as SD's text condition using CLIP image features from LQ inputs. A cross-modal adapter (CM-Adapter) converts CLIP image embeddings into degradation-aware text vectors, dynamically integrated into HQ textual embeddings based on degradation probabilities estimated by a lightweight predictor. For visual prompts, MPerceiver acquires multiscale detail representations that are crucial for image restoration. An image restoration adapter (IR-Adapter) decomposes VAE image embeddings into multi-scale features, dynamically modulated by visual prompts. This dynamic integration in both textual and visual prompt learning enables adaptation to diverse degradations and improves generalization by treating training-unseen degradations as a combination of those in the training set. Additionally, a detail refinement module (DRM) extracts degradation-aware LQ features from the VAE encoder, fused into the decoder through direct encoder-to-decoder information transformation, further enhancing fidelity.

To demonstrate the superiority of MPerceiver as an all-in-one approach, it's trained on 9 IR tasks covering both synthetic and real settings. As shown in Fig. 2, our method outperforms all compared all-in-one methods and achieves even better results than state-of-the-art task-specific methods in many tasks. Besides, MPerceiver can even handle challenging mixed degradations that may occur in real-world scenarios (Fig. 1 I). Furthermore, after multitask pre-training, MPerceiver has learned general representations in low-level vision. Comprehensive experiments show that pre-trained MPerceiver exhibits favorable zero-shot and few-shot capabilities in 6 unseen tasks (Fig. 1 II, III).

The main contributions can be summarized as follows:

- We propose a novel multimodal prompt learning approach to fully exploit the generative priors of Stable Diffusion for better adaptiveness, generalizability and fidelity of all-in-one image restoration.

- We propose a dual-branch module with CM-Adapter and IR-Adapter to learn holistic and multiscale detail representations, respectively. The dynamic integration mechanism for textual and visual prompts enables adaptation to diverse, unknown degradations.
- Extensive experiments on 16 IR tasks (all-in-one, zero-shot, few-shot) validate MPerceiver's superiority in achieving high adaptiveness, robust generalizability, and superior fidelity when addressing intricate degradations.

2. Related Work

2.1. Image Restoration

Image restoration (IR) methods for known degradations [4, 10, 11, 57, 58, 79, 109, 121, 122, 126, 135, 139, 140] have been widely explored, while all-in-one approaches are still in the exploratory stage [13, 54, 63, 82]. TransWeather [104] designs a transformer-based network with learnable weather type queries to tackle different types of weather. AirNet [49] recovers various degraded images through a contrastive-based degraded encoder. Zhu *et al*. [142] propose a strategy for investigating both weather-general and weather-specific features. IDR [124] employs an ingredients-oriented paradigm to investigate the correlation among various restoration tasks. Most of the existing methods are capable of handling a limited range of degradation types and cannot cover complex real-world scenarios.

Since diffusion models have shown a strong capability to generate realistic images [19, 35, 90, 96, 101], several diffusion-based methods have been proposed for image restoration [55]. These methods can primarily be categorized into zero-shot and supervised learning-based approaches. Zero-shot methods [15, 16, 21, 46, 100, 108, 143] leverage pre-trained diffusion models as generative priors, seamlessly incorporating degraded images as conditions into the sampling process. Supervised learning-based methods [52, 76, 81, 95, 97, 112] train a conditional diffusion model from scratch. Recently, several approaches [61, 106] have endeavored to employ pre-trained text-to-image diffusion models for blind image super-resolution.

2.2. Prompt Learning for Vision Tasks

Inspired by the success of prompt learning in NLP [8, 25, 56, 65], the computer vision community has begun to explore its applicability to vision [40, 43, 64, 110] and vision-language models [22, 91, 137, 138]. CoOp [138] learns a set of task-specific textual prompts to fine-tune CLIP [89] for downstream image recognition. CoCoOP [137] refines the generalizability of CoOp by learning a lightweight neural network to generate image-conditional dynamic prompts. VPT [43] learns a set of visual prompts to finetune transformer-based vision models for downstream recognition tasks. Compared to concurrent works that introduce prompt learning into IR [66, 70, 71, 83], ours is the first to explore multimodal prompt design in low-level vision.

3. Method

We propose MPerceiver for all-in-one image restoration in complex real-world scenarios. First, we review the latent diffusion models [93] in Sec. 3.1. To effectively leverage priors in SD, we propose a dual-branch module with the cross-modal adapter (CM-Adapter) and image restoration adapter (IR-Adapter) and encode degradation-dedicated information into multimodal prompts, which is illustrated in Sec. 3.2. Finally, we introduce a detail refinement module (DRM) to enhance the restoration fidelity in Sec. 3.3.

3.1. Preliminary: Latent Diffusion Models

Our method is based on Stable Diffusion (SD) [93], a text-to-image diffusion model that conducts the diffusion-denoising process in the latent space. SD utilizes a pre-trained VAE to encode images into latent embeddings z_0 and then trains the denoising U-Net ϵ_θ in the latent space, which can be formulated as

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,c,t,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t)||_2^2], \quad (1)$$

where $\epsilon \in \mathcal{N}(0, \mathbf{I})$ is the ground truth noise map at time step t. c represents the conditional information. $\bar{\alpha}_t$ is the diffusion coefficient in DDPM [35].

3.2. Dual-branch with Multimodal Prompts

As shown in Fig. 3, the proposed MPerceiver adopts a dual-branch (*i.e.*, textual branch and visual branch) module with textual and visual prompts in their corresponding branch. Motivated by CLIP's powerful representation capability for images [26, 33, 75], we utilize the pre-trained image encoder $\mathcal{E}_{clip}^{img}$ to extract features rich in degradation-aware information. The degraded features e_{clip}^{img} will be fed into a lightweight trainable degradation predictor to provide predictions $P \in \mathbb{R}^N$ (N denotes the number of degradations), which will serve as dynamic weights to adjust the integration process of multimodal prompts. The degradation predictor is optimized through focal loss [60].

Textual branch with CM-Adapter. To effectively leverage the powerful text-to-image generation capability of SD, we aim to obtain the text description of desired HQ images. As shown in Fig. 3, we propose a cross-modal adapter (CM-Adapter) with the cross-modal inversion mechanism to transform CLIP LQ image embeddings e_{clip}^{img} to desired HQ text embeddings e_{clip}^{txt}.

Specifically, LQ image embeddings e_{clip}^{img} will first go through a small network for degradation-agnostic mapping. Then we employ a series of parallel self-attention [105] layers as the degradation-dedicated mapping to obtain one set

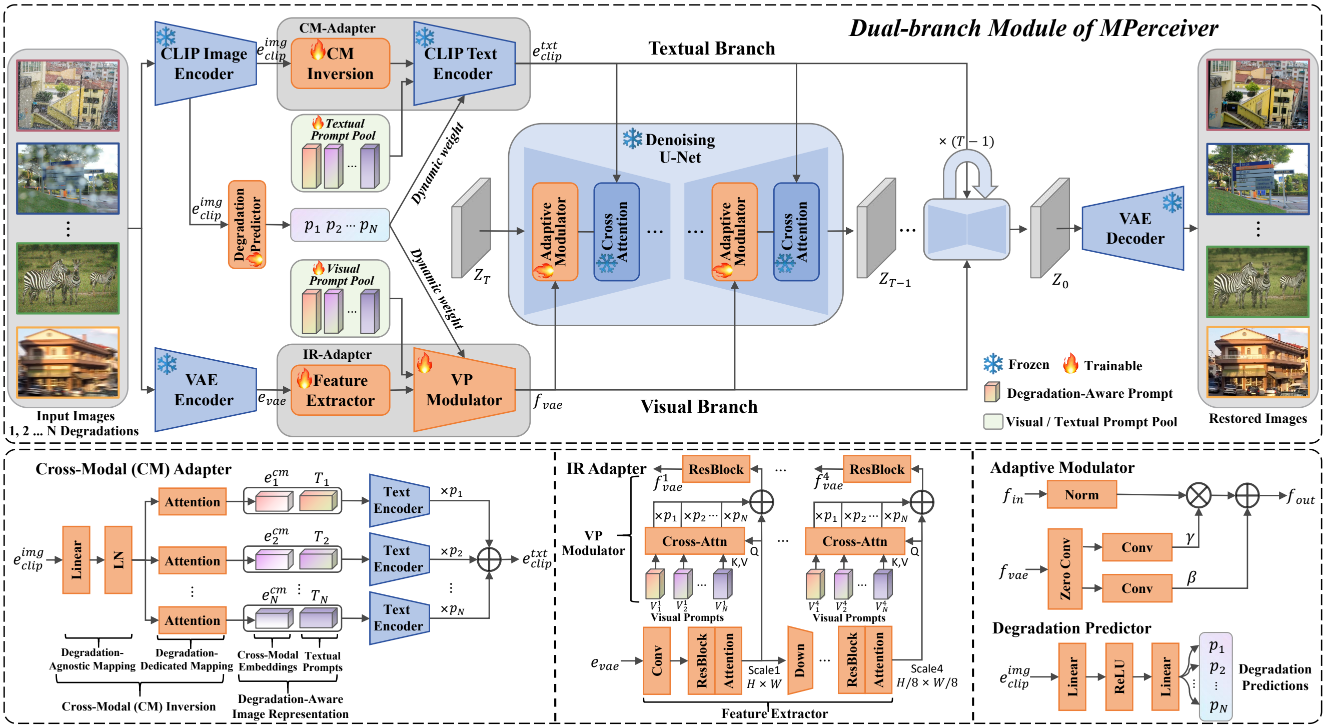


Figure 3. Illustration of MPerceiver's dual-branch module with multimodal prompts. *Textual Branch*: CLIP image embeddings are transformed into text vectors through cross-modal inversion, which are then used alongside textual prompts as holistic representations for SD. *Visual Branch*: IR-Adapter decomposes VAE image embeddings into multi-scale features, which are then dynamically modulated by visual prompts to provide detail guidance for SD adaptively.

of degradation-dedicated cross-modal embeddings (i.e., text vectors) $e^{cm} = \{e_1^{cm}, \cdots, e_N^{cm}\}$, where e_i^{cm} corresponds to a specific type of degradation. The whole process of cross-modal inversion can be formulated as:

$$e_i^{cm} = \mathrm{Attn}_i(\mathrm{LN}(\mathrm{FC}(e_{clip}^{img}))), i \in \{1, \cdots, N\}, \quad (2)$$

where Attn_i is the self-attention layer for the i-th degradation, and N is the number of degradations.

Furthermore, we establish a textual prompt (TP) pool $T = \{T_1, \cdots, T_N\} \in \mathbb{R}^{N\times L\times C_{clip}}$ to encapsulate degradation-dedicated information, where L represents the number of tokens, and C_{clip} is the embedding dimension of CLIP. Serving as learnable parameters, T collaborates with cross-modal embeddings e^{cm} to constitute a comprehensive set of SD text prompts, functioning as representations aware of image degradations. Given degradation probabilities $P = \{p_1, \cdots, p_N\} \in \mathbb{R}^N$ estimated by the degradation predictor, we dynamically integrate the set of text prompts to obtain the High-Quality (HQ) text embeddings:

$$e_{clip}^{txt} = \sum_{i=1}^{N} p_i \mathcal{E}_{clip}^{txt}(\{e_i^{cm}, T_i\}). \quad (3)$$

where $\mathcal{E}_{clip}^{txt}$ denotes the CLIP text encoder. Finally, we integrate e_{clip}^{txt} into SD through a frozen cross-attention layer to provide a holistic representation.

Visual branch with IR-Adapter. The textual branch can provide a holistic representation for SD, but it lacks detailed information that is crucial for restoration fidelity. A visual branch is introduced to extract multi-scale detail representations and complement with the textual branch. As shown in Fig. 3, we first project degraded images into latent embeddings $e_{vae} \in \mathbb{R}^{H\times W\times 4}$ through the VAE encoder of SD. Then we propose an image restoration adapter (IR-Adapter) to acquire multi-scale detail features as guidance for SD. We utilize a feature extractor to decompose e_{vae} into multi-scale features $f_{vae} = \{f_{vae}^1, f_{vae}^2, f_{vae}^3, f_{vae}^4\}$. In each scale, we employ a residual block (RB) and a self-attention layer to extract features. Similar to the textual prompt pool, we construct 4 visual prompt pools $V = \{V^k, |k \in \{1,2,3,4\}\}$ for each scale, where $V^k \in \mathbb{R}^{N\times M\times C_k}$, M is a hyper-parameter specifying the capacity of visual prompt (VP) pools and C_k is the channel dimension of the k-th scale. As shown in Fig. 3, we propose a visual prompt (VP) modulator to dynamically integrate the degradation-aware information provided by visual prompts into multi-scale features, formulated as:

$$f_{vae}^k = \mathrm{RB}_k(f_{vae}^k + \sum_{i=1}^{N} p_i \mathrm{MHCA}_k(f_{vae}^k, V_i^k, V_i^k)), \quad (4)$$

where $\mathrm{MHCA}_k(q, k, v)$ is the multi-head cross-attention layer of the k-th scale. Then the degradation-aware multi-scale features will modulate the features in the U-Net with the operation of AdaIN [42], formulated as:

$$f_{out} = \gamma(f_{vae}) \odot \mathrm{Norm}(f_{in}) + \beta(f_{vae}), \tag{5}$$

where $\odot$ denotes the element-wise multiplication, f_{in} is the original feature in the U-Net, f_{out} is the feature after modulation, $\gamma(\cdot)$ and $\beta(\cdot)$ are implemented by two convolutional layers. The dual-branch module is optimized directly using the latent diffusion loss in Eq. (1).

Note that the dynamic integration mechanisms in Eq. (3) and Eq. (4) significantly augment the adaptiveness and generalizability of MPerceiver when confronting diverse degradations. In the case of degradations present during training, the model can discern their types and select corresponding textual and visual prompts for Stable Diffusion (SD). For those unseen during training (especially for mixed ones that often occur in real-world scenarios), it treats them as a probabilistic combination of known training degradations.

3.3. Detail Refinement Module

While the proposed dual-branch module empowers Stable Diffusion (SD) with a robust ability to identify and eliminate degradations in images, the generated images may exhibit a tendency to lose details of small objects. This effect is attributed to the high compression rate of the autoencoder employed by SD, as discussed in previous works [18, 145]. For instance, in Fig. 5 (b), although the degradations in the images have been largely addressed, noticeable artifacts become apparent, particularly affecting small text.

To address this concern, we introduce a Detail Refinement Module (DRM) aimed at providing supplementary information to assist the SD VAE decoder in the image reconstruction process. As depicted in Fig. 4, DRM functions as a plug-in module, enabling direct encoder-to-decoder information transformation through a skip connection. Following modulation by the visual prompt (VP) modulator, Low-Quality (LQ) features f_{lq} are concatenated with the original decoder features f_{in}. Subsequently, a sequence of ResBlocks [11] and SwinBlocks [58] is employed to extract auxiliary features, enhancing detail reconstruction before the residual connection. The training of DRM involves a combination of reconstruction (L1) loss, color loss [107], perceptual loss [45], and adversarial loss [27].

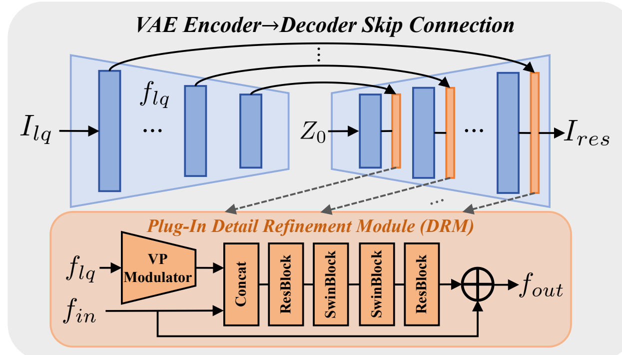


Figure 4. Illustration of the detail refinement module (DRM). For simplicity, visual prompts and degradation predictions are omitted as input to the visual prompt (VP) modulator. Apart from the DRM which is trainable, the other modules are all frozen.

(a) Input (b) w/o DRM (c) w/ DRM

Figure 5. Effect of the DRM on raindrop removal (top row from [84]) and motion deblurring (bottom row from [80]). The proposed DRM significantly improves the fidelity of the results.

4. Experiments

4.1. Experimental Setup

Settings. (1) All-in-one: We train a unified model to solve 10 IR tasks, including deraining, dehazing, desnowing, raindrop removal, low-light enhancement, motion deblurring, defocus deblurring, gaussian denoising, real denoising and challenging mixed degradations removal. **(2) Zero-shot:** We use the all-in-one pre-trained model to directly solve training-unseen tasks, including under-display camera IR (POLED/TOLED), underwater IR. **(3) Few-shot:** We fine-tune the all-in-one pre-trained model using a small amount of data (about 3%-5% of the data used by task-specific methods) and adapt it to new tasks, including JPEG compression artifact removal, demosaicking, demoireing.

Datasets and Metrics. For **setting(1)**, a combination of various image degradation datasets is used to evaluate our method, *i.e.*, Rain1400 [24], Outdoor-Rain [53], SSID [38] and LHP [31] for deraining; RESIDE [48], NH-HAZE [6] and Dense-Haze [5] for dehazing; Snow100K [68] and RealSnow [142] for desnowing; RainDrop [84] and RainDS [86] for raindrop removal; LOL-v2 [119] for low-light enhancement; CBSD68, CBSD400 [74], Urban100 [39], Kodak24 [23], McMaster [132], WED [72] and DF2K for gaussian denoising; SIDD [1] for real image denoising; GoPro [80] and RealBlur [92] for motion deblurring; and DPDD [2] for defocus deblurring. Considering real-world LQ images may contain more than just a single degradation, we construct a challenging mixed degradation benchmark named MID6, in which the LQ images contain mixed degradations (*e.g.*, low-light&noise&blur, rain&raindrop&noise; See Fig. 7). For **setting(2)**, we utilize TOLED [141] and POLED [141] for under-display camera (UDC) IR. Following [41], 90 pairs from UIEB [50] and 110 pairs from UWCNN [51] are used for underwater IR. For **setting(3)**, we use the first 100 images of DIV2K [3]

Table 1. **[All-in-one]** Quantitative comparison with state-of-the-art task-specific methods and all-in-one methods on 9 tasks. General IR models trained under the all-in-one setting are marked with symbol $(\cdot)_A$. Best and second best performance are in red and blue colors, respectively. When using the self-ensemble strategy, the model is marked with "+".

Type	Method	*Deraining* (Rain1400)		Method	*Dehazing* (Average)		Method	*Desnowing* (Snow100K-L)	
		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓
Task Specific	Uformer [109]	32.84 / 0.931	23.31 / 0.061	AECRNet [113]	17.84 / 0.546	225.8 / 0.526	DesnowNet [68]	27.17/ 0.898	- / -
	Restormer [122]	33.68 / 0.939	20.33 / 0.050	SGID [7]	14.36 / 0.562	342.9 / 0.580	DDMSNet [129]	28.85 / 0.877	3.24 / 0.096
	DRSformer [14]	33.66 / 0.939	20.06 / 0.050	DeHamer [29]	18.64 / 0.622	241.6 / 0.488	DRT [59]	29.56 / 0.892	8.15 / 0.135
	UDR-S^2 [12]	33.08 / 0.930	19.89 / 0.053	MB-Taylor [85]	17.94 / 0.602	250.8 / 0.499	WeatherDiff [81]	30.43 / 0.915	2.81 / 0.100
All in One	AirNet [49]	32.36 / 0.928	22.38 / 0.058	AirNet [49]	16.48 / 0.589	219.9 / 0.479	AirNet [49]	30.14 / 0.907	3.92 / 0.105
	PromptIR [83]	33.26 / 0.935	22.59 / 0.058	PromptIR [83]	16.97 / 0.595	231.9 / 0.471	PromptIR [83]	30.91 / 0.913	3.79 / 0.100
	DA-CLIP [70]	29.67 / 0.851	35.01 / 0.116	DA-CLIP [70]	15.01 / 0.544	224.6 / 0.468	DA-CLIP [70]	28.31 / 0.862	3.11 / 0.098
	Restormer_A [122]	33.09 / 0.933	24.14 / 0.061	Restormer_A [122]	15.86 / 0.584	221.4 / 0.477	Restormer_A [122]	30.98 / 0.914	4.54 / 0.104
	NAFNet_A [11]	33.27 / 0.936	22.39 / 0.050	NAFNet_A [11]	15.97 / 0.597	228.6 / 0.454	NAFNet_A [11]	31.42 / 0.920	2.72 / 0.091
	MPerceiver(Ours)	33.40 / 0.937	17.82 / 0.049	**MPerceiver(Ours)**	20.95 / 0.644	196.8 / 0.437	**MPerceiver(Ours)**	31.02 / 0.916	2.31 / 0.087
	MPerceiver+(Ours)	33.69 / 0.940	17.36 / 0.047	**MPerceiver+(Ours)**	21.08 / 0.651	190.1 / 0.422	**MPerceiver+(Ours)**	31.11 / 0.918	2.14 / 0.085
Type	**Method**	***Raindrop Removal* (RainDrop)**		**Method**	***Low-light Enhance.* (LOL-v2-Real)**		**Method**	***Motion Deblur* (GoPro)**	
		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓
Task Specific	AttentGAN [84]	31.59 / 0.917	33.33 / 0.056	SNR [117]	21.48 / 0.849	58.76 / 0.159	MPRNet [121]	32.66 / 0.959	10.98 / 0.091
	Quan *et al.* [87]	31.37 / 0.918	30.56 / 0.065	SNR-SKF [114]	21.93 / 0.842	73.70 / 0.160	Restormer [122]	32.92 / 0.961	10.63 / 0.086
	IDT [116]	31.87 / 0.931	25.54 / 0.059	RQ-LLIE [67]	22.37 / 0.854	56.92 / 0.143	Stripformer [103]	33.08 / 0.962	9.03 / 0.079
	UDR-S^2 [12]	32.64 / 0.942	27.17 / 0.064	Retinexformer [9]	22.80 / 0.840	62.45 / 0.169	DiffIR [115]	33.20 / 0.963	9.65 / 0.081
All in One	AirNet [49]	31.32 / 0.925	33.34 / 0.073	AirNet [49]	19.69 / 0.821	55.43 / 0.151	AirNet [49]	28.31 / 0.910	15.31 / 0.122
	PromptIR [83]	32.03 / 0.938	35.75 / 0.073	PromptIR [83]	21.23 / 0.860	53.92 / 0.145	PromptIR [83]	31.02 / 0.938	17.54 / 0.131
	DA-CLIP [70]	30.44 / 0.880	29.38 / 0.078	DA-CLIP [70]	21.76 / 0.762	48.23 / 0.134	DA-CLIP [70]	27.12 / 0.823	16.81 / 0.136
	Restormer_A [122]	31.75 / 0.936	38.22 / 0.075	Restormer_A [122]	20.77 / 0.851	57.04 / 0.155	Restormer_A [122]	30.59 / 0.934	14.56 / 0.115
	NAFNet_A [11]	32.79 / 0.943	29.80 / 0.063	NAFNet_A [11]	18.04 / 0.827	54.25 / 0.147	NAFNet_A [11]	32.01 / 0.953	13.42 / 0.101
	MPerceiver(Ours)	33.21 / 0.929	21.27 / 0.051	**MPerceiver(Ours)**	22.16 / 0.848	45.90 / 0.130	**MPerceiver(Ours)**	32.49 / 0.959	10.69 / 0.089
	MPerceiver+(Ours)	33.62 / 0.930	19.37 / 0.044	**MPerceiver+(Ours)**	22.49 / 0.854	45.29 / 0.129	**MPerceiver+(Ours)**	32.98 / 0.961	10.51 / 0.087
Type	**Method**	***Defocus Deblur* (DPDD)**		**Method**	***Gaussian Denoising* (Average)**		**Method**	***Real Denoising* (SIDD)**	
		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓
Task Specific	DRBNet [94]	25.73 / 0.791	49.04 / 0.183	SwinIR [58]	29.60 / 0.842	58.99 / 0.146	MPRNet [121]	39.71 / 0.958	49.54 / 0.200
	Restormer [122]	25.98 / 0.811	44.55 / 0.178	Restormer [122]	29.73 / 0.845	57.56 / 0.148	Uformer [109]	39.89 / 0.960	47.18 / 0.198
	NRKNet [88]	26.11 / 0.810	55.23 / 0.210	ART [126]	29.79 / 0.845	58.50 / 0.141	Restormer [122]	40.02 / 0.960	47.28 / 0.195
	FocalNet [17]	26.18 / 0.808	48.82 / 0.210	Xformer [125]	29.83 / 0.847	55.22 / 0.144	ART [126]	39.99 / 0.960	42.38 / 0.189
All in One	AirNet [49]	25.37 / 0.770	58.82 / 0.193	AirNet [49]	28.37 / 0.801	69.36 / 0.181	AirNet [49]	38.32 / 0.945	51.20 / 0.134
	PromptIR [83]	25.66 / 0.791	52.64 / 0.197	PromptIR [83]	28.82 / 0.816	63.76 / 0.170	PromptIR [83]	39.52 / 0.954	50.52 / 0.198
	DA-CLIP [70]	24.91 / 0.749	57.43 / 0.201	DA-CLIP [70]	25.13 / 0.692	59.82 / 0.235	DA-CLIP [70]	34.04 / 0.824	34.56 / 0.186
	Restormer_A [122]	25.74 / 0.795	54.74 / 0.213	Restormer_A [122]	28.65 / 0.812	63.48 / 0.172	Restormer_A [122]	39.48 / 0.954	51.75 / 0.190
	NAFNet_A [11]	25.85 / 0.803	48.45 / 0.191	NAFNet_A [11]	29.21 / 0.829	60.84 / 0.163	NAFNet_A [11]	39.76 / 0.957	45.54 / 0.197
	MPerceiver(Ours)	25.88 / 0.803	48.22 / 0.190	**MPerceiver(Ours)**	29.57 / 0.838	61.44 / 0.158	**MPerceiver(Ours)**	39.96 / 0.959	41.11 / 0.191
	MPerceiver+(Ours)	26.06 / 0.805	46.07 / 0.190	**MPerceiver+(Ours)**	29.61 / 0.839	60.91 / 0.156	**MPerceiver+(Ours)**	40.05 / 0.960	41.46 / 0.190

Table 2. **[All-in-one]** Quantitative comparison on the proposed mixed degradation benchmark MID6.

Method	Haze &Noise &Blur			Lowlight &Noise &Blur			Rain &Noise &Blur			Rain &Raindrop &Noise			Raindrop &Noise &Blur			Snow &Noise &Blur		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
AirNet [49]	17.51	0.613	0.440	17.09	0.552	0.527	24.31	0.601	0.364	21.46	0.570	0.471	26.44	0.745	0.353	22.62	0.566	0.454
TransWeather [104]	25.11	0.739	0.241	20.36	0.626	0.408	25.01	0.683	0.294	21.59	0.541	0.412	27.76	0.737	0.229	23.90	0.672	0.306
WGWS-Net [142]	17.66	0.617	0.394	17.57	0.570	0.448	22.10	0.600	0.353	20.12	0.522	0.446	25.67	0.718	0.307	20.03	0.569	0.401
PromptIR [83]	18.41	0.631	0.437	20.95	0.649	0.413	23.75	0.647	0.313	21.31	0.556	0.461	25.41	0.721	0.327	20.92	0.601	0.362
Restormer_A [122]	17.03	0.602	0.470	16.49	0.541	0.498	23.22	0.611	0.332	20.39	0.561	0.493	24.48	0.697	0.401	21.39	0.606	0.392
NAFNet_A [11]	16.59	0.548	0.541	15.72	0.605	0.520	23.48	0.563	0.346	22.72	0.599	0.439	27.20	0.769	0.307	20.28	0.535	0.484
MPerceiver (Ours)	26.19	0.782	0.211	23.84	0.671	0.343	26.00	0.762	0.193	22.35	0.525	0.268	28.49	0.771	0.127	24.36	0.719	0.263

to fine-tune our method for JPEG compression artifact removal and demosaicking. 5% of the data in TIP2018 [102] training set is used to fine-tune our method for demoireing. We provide a detailed introduction to the datasets used for training and testing in the Appendix.

We adopt PSNR and SSIM as the distortion metrics, LPIPS [133] and FID [34] as the perceptual metrics, NIQE [78] and BRISQUE [77] as no-reference metrics.

Implementation Details. We use the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with the initial learning rate $1e^{-4}$ gradually reduced to $1e^{-6}$ with cosine annealing [69] to train our model. The training runs for 500 epochs with

Table 3. **[All-in-one]** Quantitative comparison on real-world datasets of ***deraining***, ***desnowing*** and ***motion deblurring***. SSID [38] has no GT images. Methods are directly applied to the LHP [31] and RealBlur-J [92] sets to evaluate generalization to real-world images. MUSS [38] is a semi-supervised deraining model trained with additional real rainy data, denoted with * for reference.

Type	*Deraining* (LHP [31])			*Deraining* (SSID [38])			*Desnowing* (RealSnow [142])			*Motion Deblur* (RealBlur-J [92])		
	Method	PSNR ↑	SSIM ↑	Method	NIQE ↓	BRISQUE ↓	Method	PSNR ↑	SSIM ↑	Method	PSNR ↑	SSIM ↑
Task Specific	MUSS* [38]	30.02	0.886	MUSS* [38]	3.43	28.97	MIRNetv2 [123]	31.39	0.916	MPRNet [121]	28.70	0.873
	Restormer [122]	29.72	0.889	Restormer [122]	4.12	33.29	ART [126]	31.05	0.913	Restormer [122]	28.96	0.879
	DRSformer [14]	30.04	0.895	DRSformer [14]	4.19	35.52	Restormer [122]	31.38	0.923	Stripformer [103]	28.82	0.876
	UDR-S^2 [12]	28.59	0.884	UDR-S^2 [12]	3.77	35.86	NAFNet [11]	31.44	0.919	DiffIR [115]	29.06	0.882
All in One	AirNet [49]	31.73	0.889	AirNet [49]	3.69	30.91	AirNet [49]	31.02	0.923	AirNet [49]	27.91	0.834
	TransWeather [104]	29.87	0.867	TransWeather [104]	3.96	30.94	TransWeather [104]	31.13	0.922	TransWeather [104]	28.03	0.837
	WGWS-Net [142]	30.77	0.885	WGWS-Net [142]	3.71	30.79	WGWS-Net [142]	31.37	0.919	WGWS-Net [142]	28.10	0.838
	MPerceiver (Ours)	32.07	0.889	**MPerceiver** (Ours)	3.60	30.77	**MPerceiver** (Ours)	31.45	0.924	**MPerceiver** (Ours)	29.13	0.881

Table 4. **[Zero-shot]** ***UDC IR*** (TOLED / POLED) results.

Method	TOLED [141]			POLED [141]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
AirNet [49]	26.76	0.799	0.307	13.49	0.522	0.696
TransWeather [104]	27.58	0.810	0.316	15.86	0.590	0.707
WGWS-Net [142]	22.11	0.731	0.374	10.96	0.429	0.776
Restormer$_A$ [122]	27.74	0.841	0.294	13.94	0.528	0.681
NAFNet$_A$ [11]	27.90	0.848	0.320	10.68	0.555	0.713
MPerceiver (Ours)	32.92	0.863	0.161	20.41	0.650	0.445

Table 5. **[Zero-shot]** ***Underwater IR*** results.

Method	UIEB [50]			UWCNN[51]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
AirNet [49]	17.09	0.761	0.304	13.54	0.737	0.403
TransWeather [104]	17.17	0.754	0.303	13.59	0.731	0.408
WGWS-Net [142]	16.99	0.745	0.340	13.83	0.740	0.410
Restormer$_A$ [122]	17.34	0.770	0.300	13.49	0.737	0.401
NAFNet$_A$ [11]	17.31	0.736	0.307	13.62	0.736	0.405
MPerceiver (Ours)	22.69	0.902	0.150	14.77	0.774	0.299

Table 6. **[Few-shot]** ***Color JPEG compression artifact removal*** (QF=10) results. We only use 100 images from DIV2K to fine-tune all-in-one methods, while task-specific methods adopt DIV2K and Flickr2K as the training set (3450 images).

Type	Method	LIVE1 [98]		BSD500 [74]	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Task Specific	QGAC [20]	27.62	0.804	27.74	0.802
	FBCNN [44]	27.77	0.803	27.85	0.799
All-in-One	AirNet [49]	27.47	0.797	27.60	0.788
	TransWeather [104]	26.45	0.755	26.68	0.785
	WGWS-Net [142]	26.50	0.750	26.60	0.741
	MPerceiver (Ours)	27.79	0.804	27.88	0.795

Table 7. **[Few-shot]** ***Image demosaicking*** results. The training setting is the same as JPEG compression artifact removal.

Datasets	RLDD [30]	RNAN [134]	DRUNet [130]	AirNet [49]	TransWeather [104]	WGWS-Net [142]	**MPerceiver** (Ours)
Kodak [23]	42.49	43.16	42.68	40.55	39.58	41.22	43.06
McMaster [132]	39.25	39.70	39.39	37.36	36.68	38.06	39.68

Table 8. **[Few-shot]** Quantitative comparison on the ***demoireing*** dataset TIP2018 [102]. Note that we only use 5% of the training data to fine-tune pre-trained all-in-one models.

Type	Method	Venue	PSNR ↑	SSIM ↑
Task Specific	MBCNN [136]	*CVPR' 20*	30.03	0.893
	FHDe^{2}Net [32]	*ECCV' 20*	27.78	0.896
	WDNet [62]	*ECCV' 20*	28.08	0.904
	ESDNet [120]	*ECCV' 22*	30.11	0.920
All-in-One	AirNet [49]	*CVPR' 22*	28.59	0.866
	TransWeather [104]	*CVPR' 22*	27.68	0.848
	WGWS-Net [142]	*CVPR' 23*	28.13	0.861
	MPerceiver (Ours)	-	30.19	0.885

512×512 patches on 8 NVIDIA A100 GPUs. We adopt DDIM [99] as our sampling strategy (50 steps). Our model is based on SD 2.1. More details are presented in Appendix.

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

All-in-one. We conduct comparisons between our method and SOTA all-in-one methods as well as task-specific methods. To ensure a fair evaluation, we train all-in-one models from scratch employing our training strategy. Given that Restormer [122] and NAFNet [11] serve as strong general IR baselines, we additionally train both a Restormer and a NAFNet model within the all-in-one setting.

Table 1 illustrates comprehensive performance comparisons with SOTA methods across 9 tasks. Our method consistently outperforms the compared all-in-one methods on all datasets. Notably, as an all-in-one approach, our method even achieves superior results compared to other task-specific methods in many tasks. Additionally, Table 2 presents a comparison on the proposed MID6 benchmark, where our method demonstrates significant advantages in addressing challenging mixed-degraded images. Visual results of some intricate cases from MID6 are presented in Fig. 7, showcasing the enhanced effectiveness of our method in handling mixed degradations. Recognizing the significance of real-world IR challenges, we present more results on real-world datasets in Table 3. Fig. 6 offers visual comparisons across various tasks in real-world scenarios, showcasing the superiority of our method in addressing complex authentic degradations.

Zero-shot. As shown in Tables 4&5, we evaluate the performance of each all-in-one method on training-unseen tasks. MPerceiver outperforms compared methods in all metrics, demonstrating the generalizability of our approach.

Few-shot. As depicted in Tables 6&7&8, we fine-tune the all-in-one methods with limited data to tailor them for new tasks. Notably, our method achieves comparable or superior results compared to task-specific methods trained with substantial amounts of data. This underscores the efficacy of MPerceiver, demonstrating that, post multitask pretraining,

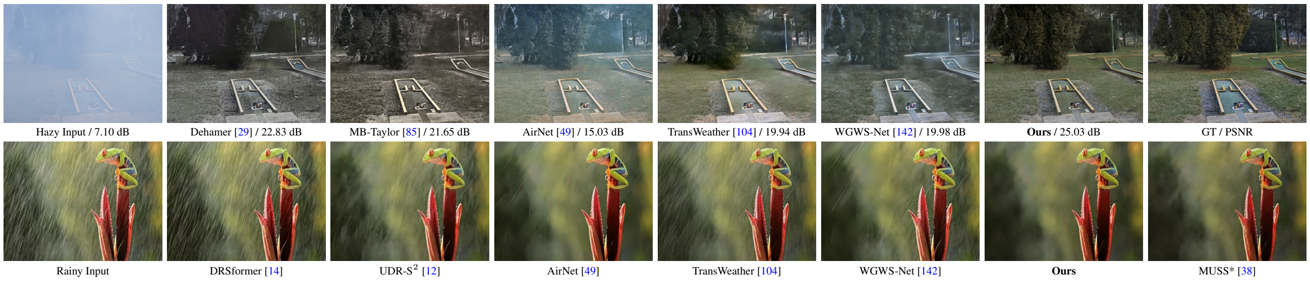
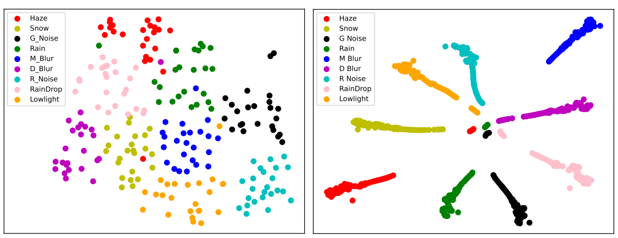


Figure 6. Real-world visual results on ***dehazing*** and ***deraining***. Best viewed with zoom in.

Figure 7. Visual results on the MID6 benchmark (R: Rain; RD: RainDrop; N: Noise; LL: Low-Light; B: Blur). Our method can better handle these challenging cases where LQ images are affected by mixed degradations compared with other all-in-one methods.



(a) t-SNE from visual prompt (b) t-SNE from CLIP textual embeddings

Figure 8. t-SNE visualizations of visual prompt V^1 and CLIP textual embeddings $\mathcal{E}_{clip}^{txt}(T)$.

Table 9. Ablations of MPerceiver. The metrics are reported on the average of deraining, dehazing and raindrop removal.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline (Stable Diffusion)	16.49	0.481	0.538
+Textual Branch	19.19	0.557	0.447
+Textual Branch w/o TP Pool	18.94	0.553	0.457
+Visual Branch	25.05	0.763	0.213
+Visual Branch w/o VP Pool	24.94	0.760	0.216
+(Visual & Textual) Branch	25.31	0.770	0.199
+(Visual & Textual) Branch w/o TP Pool	25.20	0.768	0.206
+(Visual & Textual) Branch w/o VP Pool	25.13	0.767	0.204
+(Visual & Textual) Branch + DRM (Full Model)	29.17	0.842	0.162

it has acquired general representations in low-level vision, allowing for cost-effective adaptation to new tasks.

4.3. Ablation Study

We perform ablation studies to examine the role of each component in MPerceiver. In Table 9, we initiate with SD and systematically incorporate or exclude the remaining modules of MPerceiver, including the visual branch, visual prompt (VP) pool, textual branch, textual prompt (TP) pool, and detail refinement module (DRM). The results exhibit a gradual improvement upon the addition of each component and a corresponding decline upon its removal, underscoring the effectiveness of each module. Besides, Fig. 8 visualizes the t-SNE statistics of visual prompt V^1 and CLIP textual embeddings $\mathcal{E}_{clip}^{txt}(T)$. It demonstrates that our multimodal prompt learning can effectively enable the network to distinguish different degradations.

5. Conclusion

This paper introduces MPerceiver, a multimodal prompt learning approach utilizing Stable Diffusion priors for enhanced adaptiveness, generalizability, and fidelity in all-in-one image restoration. The novel dual-branch module, comprising the cross-modal adapter and image restoration adapter, learns holistic and multiscale detail representations. The adaptability of textual and visual prompts is dynamically tuned based on degradation predictions, enabling effective adaptation to diverse unknown degradations. Additionally, a plug-in detail refinement module enhances restoration fidelity through direct encoder-to-decoder information transformation. Across 16 image restoration tasks, including all-in-one, zero-shot, and few-shot scenarios, MPerceiver demonstrates superior adaptiveness, generalizability, and fidelity.

Acknowledgements: This research is partially funded by Youth Innovation Promotion Association CAS (Grant No. 2022132), Beijing Nova Program (20230484276), National Natural Science Foundation of China (Grant No. U21B2045, U20A20223) and CAAI Huawei MindSpore Open Fund.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, pages 1692–1700, 2018. 5

[2] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *ECCV*, pages 111–126, 2020. 5

[3] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 5

[4] Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Lei Zhang, and Ran He. Uncertainty-aware source-free adaptive image super-resolution with wavelet augmentation transformer. *arXiv preprint arXiv:2303.17783*, 2023. 3

[5] 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 *ICIP*, pages 1014–1018, 2019. 5

[6] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *CVPRW*, pages 444–445, 2020. 5

[7] Haoran Bai, Jinshan Pan, Xinguang Xiang, and Jinhui Tang. Self-guided image dehazing using progressive feature fusion. *TIP*, 31:1217–1229, 2022. 6

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901, 2020. 3

[9] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, pages 12504–12513, 2023. 6

[10] 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 *CVPR*, pages 12299–12310, 2021. 3

[11] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, pages 17–33, 2022. 3, 5, 6, 7, 8

[12] 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 *ICCV*, pages 13106–13117, 2023. 6, 7, 8

[13] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *CVPR*, pages 17653–17662, 2022. 2, 3

[14] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *CVPR*, pages 5896–5905, 2023. 2, 6, 7, 8

[15] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *NeurIPS*, pages 25683–25696, 2022. 3

[16] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023. 3

[17] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *ICCV*, pages 13001–13011, 2023. 6

[18] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2, 5

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[20] Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. Quantization guided jpeg artifact correction. In *ECCV*, pages 293–309, 2020. 7

[21] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pages 9935–9946, 2023. 3

[22] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, pages 2704–2714, 2023. 3

[23] Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 4(2), 1999. 5, 7

[24] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 3855–3863, 2017. 5

[25] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 3

[26] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, https://distill.pub/2021/multimodal-neurons/, 2021. 3

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, page 2672–2680, 2014. 5

[28] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, pages 1780–1789, 2020. 2

[29] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, pages 5812–5820, 2022. 6, 8

[30] Yu Guo, Qiyu Jin, Gabriele Facciolo, Tieyong Zeng, and Jean-Michel Morel. Residual learning for effective joint demosaicing-denoising. *arXiv preprint arXiv:2009.06205*, 2020. 7

[31] Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. From sky to the ground: A large-scale bench-

mark and simple baseline towards real rain removal. In *ICCV*, pages 12097–12107, 2023. 5, 7

[32] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Fhde^{2}net: Full high definition demoireing network. In *ECCV*, pages 713–729, 2020. 7

[33] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *ICLR*, 2022. 3

[34] 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 *NeurIPS*, pages 6626–6637, 2017. 6

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 3

[36] Huaibo Huang, Aijing Yu, Zhenhua Chai, Ran He, and Tieniu Tan. Selective wavelet attention learning for single image deraining. *IJCV*, 129(4):1282–1300, 2021. 2

[37] Huaibo Huang, Aijing Yu, and Ran He. Memory oriented transfer learning for semi-supervised image deraining. In *CVPR*, pages 7732–7741, 2021. 2

[38] Huaibo Huang, Mandi Luo, and Ran He. Memory uncertainty learning for real-world single image deraining. *TPAMI*, 45(3):3446–3460, 2022. 5, 7, 8

[39] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 5

[40] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *CVPR*, pages 10878–10887, 2023. 3

[41] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *CVPR*, pages 18145–18155, 2023. 5

[42] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 5

[43] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 3

[44] Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In *ICCV*, pages 4997–5006, 2021. 7

[45] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 5

[46] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, pages 23593–23606, 2022. 3

[47] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, pages 4770–4778, 2017. 2

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

[49] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17452–17462, 2022. 2, 3, 6, 7, 8

[50] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *TIP*, 29:4376–4389, 2019. 5, 7

[51] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *PR*, 98:107038, 2020. 5, 7

[52] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 3

[53] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *CVPR*, pages 1633–1642, 2019. 5

[54] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, pages 3175–3185, 2020. 3

[55] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023. 3

[56] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, pages 4582–4597, 2021. 3

[57] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*, pages 18278–18289, 2023. 3

[58] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 3, 5, 6

[59] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *CVPRW*, pages 589–598, 2022. 6

[60] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[61] 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. 3

[62] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Aleš Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoiréing. In *ECCV*, pages 86–102, 2020. 7

[63] Lin Liu, Lingxi Xie, Xiaopeng Zhang, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, and Qi Tian. Tape: Task-agnostic prior embedding for image restoration. In *ECCV*, pages 447–464, 2022. 2, 3

[64] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *CVPR*, pages 19434–19445, 2023. 3

[65] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 3

[66] Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. *arXiv preprint arXiv:2310.10513*, 2023. 3

[67] Yunlong Liu, Tao Huang, Weisheng Dong, Fangfang Wu, Xin Li, and Guangming Shi. Low-light image enhancement with multi-stage residue quantization and brightness-aware attention. In *ICCV*, pages 12140–12149, 2023. 6

[68] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *TIP*, 27(6):3064–3073, 2018. 2, 5, 6

[69] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[70] 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*, 2023. 3, 6

[71] 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. 3

[72] 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. *TIP*, 26(2):1004–1016, 2016. 5

[73] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *CVPR*, pages 3127–3136, 2017. 2

[74] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–423, 2001. 5, 7

[75] Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In *CVPR*, pages 16410–16419, 2022. 3

[76] Ilvr: Conditioning method for denoising diffusion probabilistic models. Palette: Image-to-image diffusion models. In *ICCV*, pages 14347–14356, 2021. 3

[77] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *TIP*, 21(12):4695–4708, 2012. 6

[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. 6

[79] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *CVPR*, pages 17399–17410, 2022. 3

[80] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 3883–3891, 2017. 5

[81] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *TPAMI*, 45(8):10346–10357, 2023. 2, 3, 6

[82] 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 *CVPR*, pages 5815–5824, 2023. 2, 3

[83] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. *arXiv preprint arXiv:2306.13090*, 2023. 3, 6

[84] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, pages 2482–2491, 2018. 5, 6

[85] 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 *ICCV*, pages 12802–12813, 2023. 6, 8

[86] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *CVPR*, pages 9147–9156, 2021. 5

[87] Yuhui Quan, Shijie Deng, Yixin Chen, and Hui Ji. Deep learning for seeing through window with raindrops. In *ICCV*, pages 2463–2471, 2019. 6

[88] Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *CVPR*, pages 5754–5763, 2023. 2, 6

[89] 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 *ICML*, pages 8748–8763, 2021. 3

[90] 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

[91] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. 3

[92] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201, 2020. 5, 7

[93] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3

[94] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *CVPR*, pages 16304–16313, 2022. 2, 6

[95] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, pages 1–10, 2022. 3

[96] 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. In *NeurIPS*, pages 36479–36494, 2022. 3

[97] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 45(4): 4713–4726, 2022. 3

[98] HR Sheikh. Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 7

[99] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 7

[100] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023. 3

[101] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[102] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *TIP*, 27(8):4160–4172, 2018. 6, 7

[103] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, pages 146–162, 2022. 2, 6, 7

[104] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*, pages 2353–2363, 2022. 2, 3, 6, 7, 8

[105] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[106] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3

[107] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, pages 6849–6857, 2019. 5

[108] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023. 3

[109] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 3, 6

[110] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022. 3

[111] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *CVPR*, pages 18156–18165, 2023. 2

[112] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *CVPR*, pages 16293–16303, 2022. 3

[113] 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 *CVPR*, pages 10551–10560, 2021. 6

[114] Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, and Heng Tao Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *CVPR*, pages 1662–1671, 2023. 2, 6

[115] 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 *ICCV*, pages 13095–13105, 2023. 6, 7

[116] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *TPAMI*, 45(11): 12978–12995, 2023. 2, 6

[117] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *CVPR*, pages 17714–17724, 2022. 6

[118] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *CVPR*, pages 9893–9903, 2023. 2

[119] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *TIP*, 30:2072–2086, 2021. 5

[120] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *ECCV*, pages 646–662, 2022. 7

[121] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 3, 6, 7

[122] 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 *CVPR*, pages 5728–5739, 2022. 3, 6, 7

[123] 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. *TPAMI*, 45(2):1934–1948, 2022. 7

[124] Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *CVPR*, pages 5825–5835, 2023. 3

[125] Jiale Zhang, Yulun Zhang, Jinjin Gu, Jiahua Dong, Linghe Kong, and Xiaokang Yang. Xformer: Hybrid x-

shaped transformer for image denoising. *arXiv preprint arXiv:2303.06440*, 2023. 6

[126] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023. 3, 6, 7

[127] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7):3142–3155, 2017. 2

[128] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *TIP*, 27(9):4608–4622, 2018. 2

[129] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *TIP*, 30: 7419–7431, 2021. 6

[130] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *TPAMI*, 44(10):6360–6376, 2021. 7

[131] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023. 2

[132] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *JEI*, 20(2):023016, 2011. 5, 7

[133] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[134] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 7

[135] Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *CVPR*, pages 14122–14132, 2023. 3

[136] Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, and Ales Leonardis. Image demoireing with learnable bandpass filters. In *CVPR*, pages 3636–3645, 2020. 7

[137] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3

[138] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3

[139] Xiaoqiang Zhou, Huaibo Huang, Ran He, Zilei Wang, Jie Hu, and Tieniu Tan. Msra-sr: Image super-resolution transformer with multi-scale shared representation acquisition. In *ICCV*, pages 12665–12676, 2023. 3

[140] Xiaoqiang Zhou, Huaibo Huang, Zilei Wang, and Ran He. Ristra: Recursive image super-resolution transformer with relativistic assessment. *TMM*, pages 1–12, 2024. 3

[141] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *CVPR*, pages 9179–9188, 2021. 5, 7

[142] 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 *CVPR*, pages 21747–21758, 2023. 2, 3, 5, 6, 7, 8

[143] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *CVPRW*, pages 1219–1229, 2023. 3

[144] Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, and Shimin Hu. Traffic-sign detection and classification in the wild. In *CVPR*, pages 2110–2118, 2016. 2

[145] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion. *arXiv preprint arXiv:2306.04632*, 2023. 2, 5