Improving Image Restoration through Removing Degradations in Textual Representations

Jingbo Lin[1], Zhilu Zhang[1], Yuxiang Wei[1], Dongwei Ren[1,*], Dongsheng Jiang[2], Qi Tian[2], Wangmeng Zuo[1]
[1]Harbin Institute of Technology [2]Huawei Cloud Computing Co., Ltd.
jblincs1996@gmail.com, cszlzhang@outlook.com, yuxiang.wei.cs@gmail.com,
rendongweihit@gmail.com, dongsheng_jiang@outlook.com, tian.qi1@huawei.com, cswmzuo@gmail.com

Abstract

In this paper, we introduce a new perspective for improving image restoration by removing degradation in the textual representations of a given degraded image. Intuitively, restoration is much easier on text modality than image one. For example, it can be easily conducted by removing degradation-related words while keeping the content-aware words. Hence, we combine the advantages of images in detail description and ones of text in degradation removal to perform restoration. To address the cross-modal assistance, we propose to map the degraded images into textual representations for removing the degradations, and then convert the restored textual representations into a guidance image for assisting image restoration. In particular, We ingeniously embed an image-to-text mapper and text restoration module into CLIP-equipped text-to-image models to generate the guidance. Then, we adopt a simple coarse-to-fine approach to dynamically inject multi-scale information from guidance to image restoration networks. Extensive experiments are conducted on various image restoration tasks, including deblurring, dehazing, deraining, and denoising, and all-in-one image restoration. The results showcase that our method outperforms state-of-the-art ones across all these tasks. The codes and models are available at https://github.com/mrluin/TextualDegRemoval.

1. Introduction

Image restoration aims to reconstruct a high-quality clean image from its degraded observations. Most existing methods design deep networks for specific restoration tasks, including image denoising [55, 111, 115, 117, 119], deblurring [13, 14, 18, 93, 110], deraining [16, 93, 98, 110], dehazing [8, 19, 36, 88, 90], *etc*. Recent works [15, 50, 69, 71, 114, 124] expect to explore a unified model for multiple degradations. However, the severely ill-posed nature of the task makes it non-trivial to separate degradations and desired image content. Especially for unified models, the potential conflicts in dealing with various degradations bring more uncertainty.

From a broader perspective, the purpose of the restoration is to enhance the clarity of scenes for human perception and recognition, while performing it on image modality is just one option. Moreover, degradations in image modality are tightly coupled with desirable content, making their removal challenging. When recording the scenes in some other modalities, this problem can be alleviated. Let us take the text modality as an example. Assume the textual description of a clean scene is represented by 'a scene of $*$'. When this scene is subjected to rainfall, the description can be converted into 'a rainy scene of $*$'. Thus, deraining can be readily achieved by simply removing the rain-related text 'rainy'. Furthermore, it is also convenient to remove multiple degradations in textual space with one unified model. Note that text modality may only describe rough aspects and ignore some details. We can further introduce image modality to combine their complementary potential in degradation removal and image restoration.

Motivated by the above, we propose to perform restoration first in a modality where degradations and content are loosely coupled, and then utilize the corrected content to guide image restoration. In particular, we adopt the commonly used textual modality. Note that the text corresponding to the scene is not always available and the cross-modal assistance is difficult to achieve directly. The gap between text and image should be bridged. First, degraded images should be mapped into textual space for removing degradations. Second, the restored text should be mapped into a guidance image to assist restoration of the degraded image. As shown in Fig. 1, for the former, we have tried to convert degraded images to textual captions by image-to-text (I2T) models (*e.g.*, BLIP [51, 52]), and find the captions can explicitly express both degradation types (*e.g.*, blur, rain, haze, and noise) and content information. For the latter, we can

*Correspondence author.

Figure 1. **Overview of the proposed method**. We propose to improve image restoration by performing restoration on the textual level, in which content and degradation information are loosely coupled. We first encode image concepts into textual space and then remove degradation-related information. To achieve cross-modal assistance, we employ pre-trained T2I models to generate clean guidance for the image restoration process.

utilize the text-to-image (T2I) models (*e.g*., Imagen [82]), which have demonstrated impressive image generation capability from the given texts.

Although such a scheme is conceptually feasible, converting to explicit text may lose a lot of information from images, and additional degraded-clean text pairs need to be collected to train the text restoration module. Fortunately, Contrastive Language-Image Pre-training (CLIP) [74] implicitly aligns concepts of images and text. And we can leverage the CLIP-equipped T2I models (*e.g*., Stable Diffusion [81]) to build an end-to-end framework for generating clear guidance from degraded images gracefully. Specifically, CLIP has been demonstrated as an effective I2T converter [23]. For converting degraded images into degraded textual representations, we adopt the image encoder of CLIP, and add an I2T mapper after the encoder. For restoring degraded representations, we further append a textual restoration module after I2T mapper. Then, I2T mapper and textual restoration module can be sequentially trained by employing paired data from different image restoration datasets, without additional text data.

Due to the ease of removing degradations on textual level, we train the I2T mapper and textual restoration module with multiple degradations at once. When training is done, it can serve multiple tasks, generating content-related and degradation-free guidance images from degraded images. Finally, we adopt a simple coarse-to-fine approach to dynamically inject multi-scale information from guidance to classic image restoration networks. Extensive experiments are conducted on multiple tasks, including all-in-one restoration, image deburring, dehazing, deraining, and denoising. The results show our method achieves better universally than corresponding state-of-the-art methods. Especially for all-in-one restoration, 0.5 dB PSNR improvement is obtained in comparison with PromptIR [71].

The main contributions can be summarized as follows:

- We introduce a new perspective for image restoration, *i.e*., performing restoration first in textual space where degradations and content are loosely coupled, and then utilizing the restored content to guide image restoration.
- To address the cross-modal assistance, we propose to embed an image-to-text mapper and textual restoration module into CLIP-equipped text-to-image models to generate clear guidance from degraded images.
- Extensive experiments on multiple tasks demonstrate that our method improves the performance of state-of-the-art image restoration networks.

2. Related Work

2.1. Image Restoration

Image Restoration for Specific Tasks. Image restoration is a fundamental computer vision problem, which aims to reconstruct high-quality images from corresponding degraded inputs. In recent years, deep learning has achieved great progress in image restoration. Starting from some simple convolutional neural networks (CNNs) [24, 115], the introduction of channel-attention [121], spatial-attention [35, 109], non-local operation [57, 122], skip-connection architectures [54, 109] and multi-stage scheme [93, 110] enables image restoration performance continuously improve. With the emergence of vision transformers [26, 60], the capability of capturing long-range dependencies in the image allows transformer-based

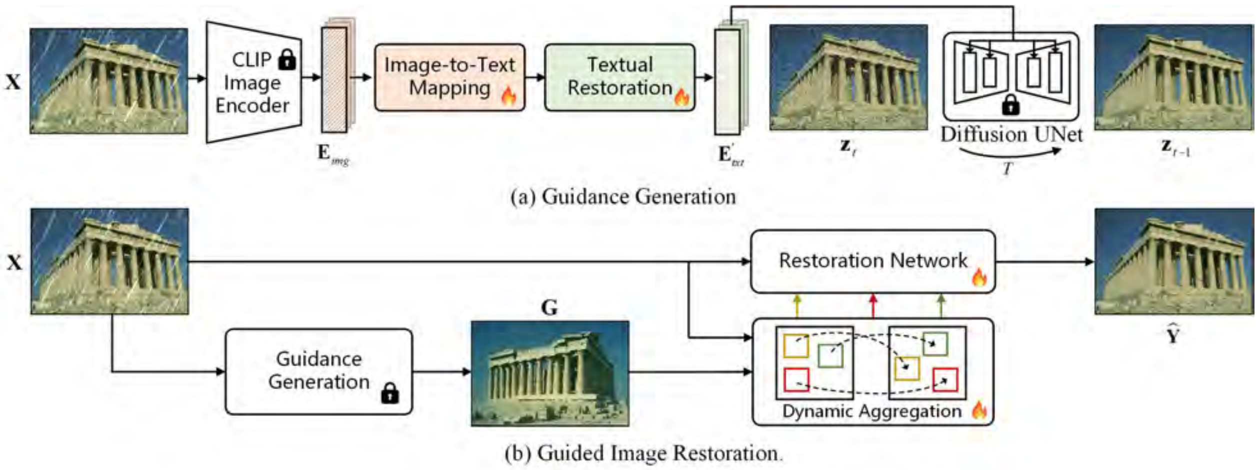


Figure 2. **Illustration of the proposed pipeline.** **(a)** We sequentially train image-to-text mapper $\mathcal{M}_{i2t}$ and textual restoration module $\mathcal{M}_{clean}$ to convert image concepts into textual representations and remove textual degradation information, respectively. **(b)** The guidance image is used to assist the image restoration process.

methods to achieve better performance, gradually replacing previous CNN-based methods. To balance computational cost, window-based attention [17, 55] and transposed attention [111] are also introduced into image restoration tasks. Albeit image restoration performance has benefited from various advanced architecture designs, most of the works only focus on one specific degradation, due to the difficulty of learning to remove multiple degradations.

All-in-One Image Restoration. Promoting by the development of unified model, some works focus on the solution to the all-in-one image restoration [15, 50, 69, 71, 114, 124]. The first all-in-one image restoration work, *i.e.*, IPT [12], employs ViT-based backbone with multi-heads and multi-tails for inputs with different degradations. However, the method is limited in handling specific synthesized degradations and cannot directly generalize to unknown tasks. AirNet [50] develops the unified model for denoising-deraining-dehazing, which utilizes contrastive learning to capture degradation representations of different tasks and adaptively injects the degradation priors into backbone restoration framework for aiding in learning better results. ADMS [69] exploits FAIG [99] in all-in-one image restoration, by learning specific filters and degradation classifiers, achieving better performance on deraining-denoising-deblurring and deraining-desnowing-dehazing tasks. IDR [114] proposes a two-stage training strategy, which first learns separate task-oriented hubs for each degradation. Then, in the second stage, it reformulates the learned hubs into a single ingredient-oriented hub by learnable PCA and adopts the reformulated hub as prior to adaptively aid in restoring corrupted images. PromptIR [71] learns to encode and adopt degradation information as prompts for dehazing-deraining-denoising task. Although existing works can unify a set of image restoration tasks into one unified model, the performance is still limited caused by the large gap among different degradations.

Prior-Based Image Restoration. Prior-based restoration aims at improving performance by introducing external priors, including structures, images, pre-trained models, *etc*. For instance, some works [9, 42, 61] adopt information from the high-resolution reference image to help improve the performance of super-resolution. Recent works [56, 92, 100] utilize pre-trained generative priors to restore more realistic and natural results. TextIR [6] develops a text-driven image restoration framework by incorporating information on textual representation.

In this work, we provide a new perspective, *i.e.*, utilizing restoration in textual space to assist image restoration. We also combine the advantages of both text-based prior (by learning degradations in textual level) and image-based prior (by giving clean guidance) for better performance.

2.2. Text-to-Image Generation

Text-to-image generation models [22, 31, 67, 75, 76, 83, 96] have attracted intensive attention in recent years due to its ability to generate high-quality and diverse images based on given text descriptions. A variety of techniques, including generative adversarial networks (GAN) [34], autoregressive models, and diffusion models have been investigated. Initial studies [48, 83, 96] mainly rely on GAN-based architectures, and train a conditional model from given paired image-caption datasets to generate samples. Some efforts focus on autoregressive models [10, 21, 31, 75, 104] have also shown exciting results. These models, such as CogView [21] and Muse [10] first learn a discrete codebook through training an autoencoder, and then adopt an autoregressive transformer to predict the tokens sequentially. With the development of diffusion models [38, 87], text conditioned image synthesis has shown remarkable improvement. By training with huge corpora, large diffusion models, such as DALLE-2 [76], Imagen [82], Stable Diffusion [80], and DALLE-3 [7] have demonstrated excellent semantic understanding, and can generate diverse and photo-realistic images.

3. Proposed Method

In this section, we present our proposed method for image restoration. As illustrated in Fig. 2, we suggest first conducting restoration in the textual modality, in which degradation is loosely coupled with the content and can be easily removed. Then we in turn utilize the restored results as guidance to improve image restoration. Specifically, in Sec. 3.2, we propose an image-to-text mapper $\mathcal{M}_{i2t}$ and a textual restoration module $\mathcal{M}_{clean}$ to extract the degradation-free textual representations $\mathbf{E}'_{txt}$ from degraded images $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$. Then with a pre-trained T2I model, we can generate a guidance image $\mathbf{G} \in \mathbb{R}^{H \times W \times 3}$ that is related to the content of $\mathbf{X}$ but free of degradation. In Sec. 3.3, to leverage clean content information from $\mathbf{G}$ for enhancing image restoration performance, we introduce a coarse-to-fine approach that dynamically incorporates multi-scale information from guidance $\mathbf{G}$ into restoration frameworks. Besides, we first give a preliminary knowledge of the T2I model in Sec. 3.1.

3.1. Preliminary

In this work, we employ Stable Diffusion [81] as our text-to-image model. Stable Diffusion pretrains an autoencoder $(\mathcal{E}(\cdot), \mathcal{D}(\cdot))$ to map the input image $\mathbf{X}$ into a lower dimensional latent space by $\mathbf{z} = \mathcal{E}(\mathbf{X})$. The decoder $\mathcal{D}(\cdot)$ learns to map the latent code back to the image as $\mathcal{D}(\mathcal{E}(\mathbf{X})) \approx \mathbf{X}$. Then, the conditional diffusion model $\epsilon_\theta(\cdot)$ is trained on the latent space to generate latent codes based on text condition $\mathbf{p}$. To train the diffusion model, simple mean-squared loss is adopted as,

$$L_{LDM} = \mathbb{E}_{\mathbf{z} \sim \mathcal{E}(\mathbf{X}), \mathbf{p}, \epsilon \sim \mathcal{N}(0,1), t} \left[\| \epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta^t(\mathbf{p})) \|_2^2 \right], \tag{1}$$

where ϵ denotes the unscaled noise, t is the timestep, $\mathbf{z}_t$ is the latent noised to time t, and $\tau_\theta^t(\cdot)$ represents the pre-trained CLIP [74] text encoder. During inference, a random Gaussian noise $\mathbf{z}_T$ is iteratively denoised to $\mathbf{z}_0$, and the final image is obtained through the decoder $\mathbf{X}' = \mathcal{D}(\mathbf{z}_0)$.

3.2. Degradation-Free Guidance Generation

Image captioning models (*e.g.*, BLIP2 [52]) can generate text descriptions of degraded images, which can be taken into the text restoration module and text-to-image models to reconstruct clean guidance images. However, such descriptions may lose too many details, resulting in severely inconsistent content between the guidance and the input images. To address these problems, we propose an image-to-text mapper $\mathcal{M}_{i2t}$ to encode the degraded image $\mathbf{X}$ as implicit textual representations $\mathbf{E}_{txt}$ rather than explicit texts, which can be used to reconstruct the content more faithfully. Then, we utilize a textual restoration module $\mathcal{M}_{clean}$ to remove the degradation from $\mathbf{E}_{txt}$ and obtain degradation-free textual representations $\mathbf{E}'_{txt}$.

Image-to-Text Mapping. Following [94], we adopt the textual word embedding space of CLIP [74] as the target space in image-to-text mapping. Benefiting from the aligned vision-language feature space of CLIP, we first adopt pre-trained CLIP image encoder $\tau_\theta^i(\cdot)$ to encode input $\mathbf{X}$ into CLIP image embedding $\mathbf{E}_{img}$. Then, we propose an image-to-text mapper $\mathcal{M}_{i2t}$ to project CLIP image embedding into textual word embedding $\mathbf{E}_{txt}$,

$$\mathbf{E}_{txt} = \mathcal{M}_{i2t}(\tau_\theta^i(\mathbf{X})), \tag{2}$$

where $\mathbf{E}_{txt} \in \mathbb{R}^{N \times D}$ and D is the dimension of textual word embedding. N is the number of learned words. To make the obtained word embedding describe more details of the input image, we set N to a large number (*e.g.*, N=20).

Textual Restoration. Although $\mathcal{M}_{i2t}$ can project images into the textual word embedding space, the projected textual word embedding will include corresponding degradation information if input $\mathbf{X}$ is degraded. Condition on the degraded textual word embedding, the synthesized guidance inevitably reflects the corresponding degradation pattern. Therefore, we further propose a textual restoration module $\mathcal{M}_{clean}$ to remove the degradation information from the textual word embedding $\mathbf{E}_{txt}$,

$$\mathbf{E}'_{txt} = \mathcal{M}_{clean}(\mathbf{E}_{txt}), \tag{3}$$

where $\mathbf{E}'_{txt}$ denote the restored textual representations. When feeding $\mathbf{E}'_{txt}$ into Stable Diffusion, we can obtain a degradation-free image $\mathbf{G}$, which has similar content as $\mathbf{X}$ but is free of degradation, as shown in Fig. 1.

Model Optimization. During training, the image-to-text mapper $\mathcal{M}_{i2t}$ and textual restoration module $\mathcal{M}_{clean}$ are optimized sequentially. In the first training stage, we collect both clean and degraded images from different restoration tasks as training data. We adopt Eq. (1) as a loss function, which constrains the diffusion model to reconstruct clean and degraded images conditioning on their own projected embedding $\mathbf{E}_{txt}$. In the second training stage, we use pairs of degraded-clean images from different image restoration datasets as training data. Based on pre-trained $\mathcal{M}_{i2t}$, we further deploy $\mathcal{M}_{clean}$ to remove degradation-related concepts from degraded embedding $\mathbf{E}_{txt}$, obtaining $\mathbf{E}'_{txt}$. We still adopt Eq. (1), which constrains the diffusion model to reconstruct clean images conditioning on restored embedding $\mathbf{E}'_{txt}$. Note that, due to the ease of removing degradations in the textual space, it is feasible to train $\mathcal{M}_{i2t}$ and $\mathcal{M}_{clean}$ with multiple degradations simultaneously. When training is done, it can serve multiple image restoration tasks, generating content-related and degradation-free guidance images $\mathbf{G}$ from degraded images

Table 1. **All-in-one image restoration results**. Following PromptIR [71], we train and evaluate the proposed method in all-in-one image restoration task, our method outperforms PromptIR across all the benchmark datasets.

Method	**Dehazing** on SOTS	**Derain** on Rain100L	**Denoise** on BSD68 $\sigma = 15$	**Denoise** on BSD68 $\sigma = 25$	**Denoise** on BSD68 $\sigma = 50$	**Average**
BRDNet [89]	23.23/0.895	27.42/0.895	32.26/0.898	29.74/0.836	26.34/0.836	27.80/0.843
LPNet [33]	20.84/0.828	24.88/0.784	26.47/0.778	24.77/0.748	21.26/0.552	23.64/0.738
FDGAN [32]	24.71/0.924	29.89/0.933	30.25/0.910	28.81/0.868	26.43/0.776	28.02/0.883
MPRNet [110]	25.28/0.954	33.57/0.954	33.54/0.927	30.89/0.880	27.56/0.779	30.17/0.899
DL [27]	26.92/0.391	32.62/0.931	33.05/0.914	30.41/0.861	26.90/0.740	29.98/0.875
AirNet [50]	27.94/0.962	34.90/0.967	33.92/0.933	31.26/0.888	28.00/0.797	31.20/0.910
PromptIR [71]	30.58/0.974	36.37/0.972	33.98/0.933	31.31/0.888	28.06/0.799	32.06/0.913
Ours	**31.63/0.980**	**37.58/0.979**	**34.01/0.933**	**31.39/0.890**	**28.18/0.802**	**32.56/0.916**

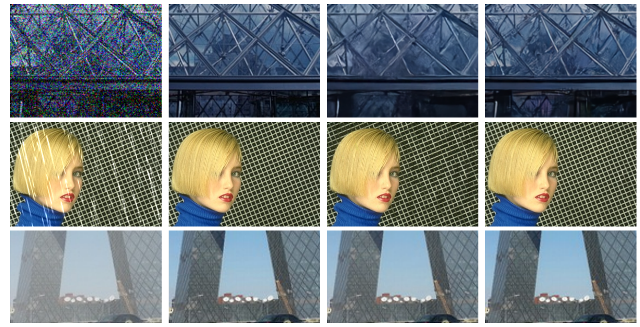


Figure 3. **All-in-one image restoration results**. Top: image denoising, mid: image deraining, bottom: image dehazing.

with one unified model (as shown in Fig. 1). More training details can be seen in the *Suppl*.

3.3. Guided Restoration

Although guidance image $\mathbf{G}$ is clean, it is inevitable that there are content differences between $\mathbf{G}$ and degraded image $\mathbf{X}$. Simply fusing the features of $\mathbf{G}$ and $\mathbf{X}$ is not enough to improve image restoration. Instead, we suggest a dynamic aggregation module to extract and exploit helpful information from $\mathbf{G}$.

Dynamic Aggregation. First, we perform feature matching [61] between the guidance $\mathbf{G}$ and given degraded image $\mathbf{X}$. Specifically, we use a shared image encoder to extract multi-scale features for $\mathbf{X}$ and $\mathbf{G}$, named $\mathbf{F}_x$ and $\mathbf{F}_g$, respectively. According to the similarity score between $\mathbf{F}_x$ and $\mathbf{F}_g$, we search for the most useful feature from $\mathbf{F}_g$ for each small patch in $\mathbf{F}_x$ in a coarse-to-fine manner. After searching, we get a set of guidance features $\hat{\mathbf{F}}_g$, whose content is aligned spatially with the input content. Second, we integrate the matched feature $\hat{\mathbf{F}}_g$ into the image restoration backbone. Both CNN-based and transformer-based restoration networks can be adopted. Without bells and whistles, we modulate the original input features as,

$$\mathbf{F}_x = \mathbf{F}_x + \alpha \cdot \mathcal{B}([\mathbf{F}_x, \hat{\mathbf{F}}_g]), \tag{4}$$

where $[\cdot, \cdot]$ represents concatenation operation, $\mathcal{B}$ represents one CNN-based block or transformer-based block, α is the hyper-parameter to trade-off feature integration. Moreover, we perform it over multiple layers or levels. More details can be seen in the *Suppl*.

Loss Function. To train guided restoration framework, we simply adopt ℓ_1 loss as the reconstruction loss between predicted results $\hat{\mathbf{Y}}$ and target $\mathbf{Y}$, *i.e.*,

$$\ell_1 = ||\hat{\mathbf{Y}} - \mathbf{Y}||_1. \tag{5}$$

4. Experiments

We evaluate the proposed method on five image restoration tasks, including **(1)** all-in-one image restoration [71], **(2)** image deblurring [14, 111] (*i.e.*, motion image deblurring and defocus image deblurring), **(3)** image dehazing [19], **(4)** image deraining [16], and **(5)** image denoising [111] (*i.e.*, grayscale and color image denoising on Gaussian noise, and real-world image denoising).

Implementation Details. We adopt publicly released stable-diffusion-v2.1-base as our T2I generative model. N is set to 20 in image-to-text mapper. Moreover, we employ the state-of-the-art methods as our image restoration backbone in different image restoration tasks: **(1)** we adopt PromptIR [71] in all-in-one image restoration, **(2)** we adopt NAFNet [14] and Restormer [111] in single-image motion deblurring, and Restormer [111] in defocus deblurring, **(3)** we adopt SFNet [19] in image dehazing, **(4)** we adopt DRSformer [16] in image deraining, **(5)** we adopt Restormer [111] in image denoising. The training settings of our guided restoration network are consistent with those of the corresponding backbone methods. Details of datasets, network architecture, and training hyper-parameters are provided in *Suppl*.

4.1. All-in-One Restoration Results

Following [71], we adopt BSD400 [5] and WED [62] for color image denoising, Rain100L [27] for image deraining, and SOTS [49] for image dehazing. We train and evaluate the proposed method on these three tasks in all-in-one image restoration setting. The large difference among different degradations makes unifying multiple image restoration tasks into one unified network difficult, and thus the performance of existing all-in-one image restoration methods is limited. Compared with performing all-in-one restoration in the image level, it is much easier for us to conduct all-in-one restoration in the textual level. Once trained, our method can generate clean textual representation for different image restoration tasks, and thus we can synthesize guidance images for various types of degradation. Comparison results in Table 1 show that our method outperforms PromptIR [71] across all the benchmark datasets, achieving +1.05 dB PSNR on dehazing task, +1.21 dB PSNR on deraining task, and +0.5 dB PSNR on average improvement. As shown in Fig. 3, the prediction of the proposed method

Table 2. **Motion image deblurring results**. We train models with GoPro training data. We evaluate our method on GoPro, HIDE, RealBlur benchmark datasets. PSNR and SSIM scores are calculated on RGB-channels.

Method	GoPro [66]		HIDE [84]		RealBlur-R [79]		RealBlur-J [79]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DBGAN [118]	31.10	0.942	28.94	0.915	33.78	0.909	24.93	0.745
MT-RNN [68]	31.15	0.945	29.15	0.918	35.79	0.951	28.44	0.862
DMPHN [113]	31.20	0.940	29.09	0.924	35.70	0.948	28.42	0.860
SPAIR [72]	32.06	0.953	30.29	0.931	-	-	28.81	0.875
MIMO-Unet+ [18]	32.45	0.957	29.99	0.930	35.54	0.947	27.63	0.837
IPT [12]	32.52	-	-	-	-	-	-	-
MPRNet [110]	32.66	0.959	30.96	0.939	35.99	0.952	28.70	0.873
HINet [13]	32.71	0.959	30.32	0.932	-	-	-	-
Uformer [93]	32.97	0.967	-	-	-	-	-	-
Restormer [111]	32.92	0.961	31.22	0.942	36.19	0.957	28.96	**0.879**
Ours-Restormer	33.11	0.962	31.26	0.943	**36.47**	**0.959**	**29.17**	0.875
NAFNet [14]	33.69	0.966	31.32	0.943	33.62	0.944	26.33	0.856
Ours-NAFNet	**33.97**	**0.968**	**31.57**	**0.946**	33.87	0.950	26.76	0.861

Table 3. **Defocus image deblurring results**. We train and evaluate methods on DPDD dataset [2]. S denotes single-image defocus deblurring model. D denotes dual-pixel defocus deblurring. PSNR and SSIM scores are calculated on RGB channels.

Method	Indoor Scenes		Outdoor Scenes		Combined	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
EBDB$_S$ [43]	25.77	0.772	21.25	0.599	23.45	0.683
DMENet$_S$ [45]	25.50	0.788	21.43	0.644	23.41	0.714
JNB$_S$ [85]	26.73	0.828	21.10	0.608	23.84	0.715
DPDNet$_S$ [2]	26.54	0.816	22.25	0.682	24.34	0.747
KPAC$_S$ [86]	27.97	0.852	22.62	0.701	25.22	0.774
IFAN$_S$ [46]	28.11	0.861	22.76	0.720	25.37	0.789
Restormer$_S$ [111]	28.87	0.882	23.24	0.743	25.98	0.811
Ours$_S$	**29.11**	**0.889**	**23.35**	**0.748**	**26.15**	**0.817**
DPDNet$_D$ [2]	27.48	0.849	22.90	0.726	25.13	0.786
RDPD$_D$ [3]	28.10	0.843	22.82	0.704	25.39	0.772
Uformer$_D$ [93]	28.23	0.860	23.10	0.728	25.65	0.795
IFAN$_D$ [46]	28.66	0.868	23.46	0.743	25.99	0.804
Restormer$_D$ [111]	29.48	0.895	23.97	0.773	26.66	0.833
Ours$_D$	**29.62**	**0.899**	**24.16**	**0.775**	**26.82**	**0.835**

Figure 4. **Motion image deblurring results**. Compared with others, our method can predict clearer results with clearer boundaries.

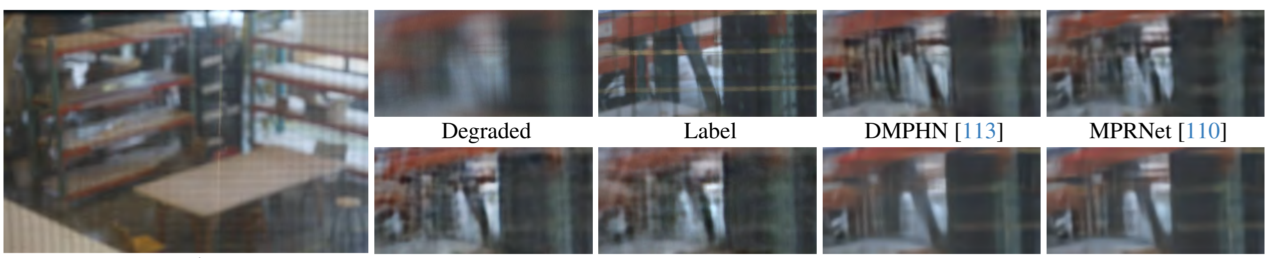


Figure 5. **Defocus image deblurring results**. Compared with others, our method can predict clearer results with clearer boundaries.

is finer and more detailed compared with PromptIR.

4.2. Image Deblurring Results

Motion Deblurring. Following [14, 111], we train the proposed method on GoPro training data and evaluate our method on GoPro [66], HIDE [84], and real-world datasets (RealBlur-R [79] and RealBlur-J [79]). As shown in Table 2, benefiting from synthesized high-quality guidance, Ours-Restormer and Ours-NAFNet outperform the state-of-the-art methods on all four benchmark datasets. Our method achieves +0.28 dB PSNR improvement on GoPro testing data. The visual results in Fig. 4 show that our method can restore images with sharper boundaries and details, demonstrating its effectiveness.

Defocus Deblurring. Following [111], we train and evaluate the proposed method on DPDD [2] dataset for single-image defocus deblurring and dual-pixel defocus deblurring. From Table 3, compared with Restormer [111], our method achieves +0.17 dB PSNR and +0.16 dB PSNR gains on single-image defocus deblurring and dual-pixel defocus deblurring, respectively. As illustrated in Fig. 5, the prediction of our method shows a clearer structure.

4.3. Image Dehazing Results

For image dehazing task, we conduct experiments on the synthetic benchmark RESIDE [49] dataset. We train the proposed method on indoor scene and outdoor scene data, and evaluate the performance on SOTS-indoor and SOTS-outdoor testing dataset [49] separately. Table 4 reports the quantitative comparison results. One can see that our method outperforms existing methods on both indoor and outdoor scenes, achieving +0.24 dB PSNR average improvement. Visual comparisons can be found in the *Suppl.*

Table 4. **Image dehazing results**. We separately train and evaluate our method indoor scene and outdoor scene. PSNR and SSIM scores are calculated on RGB-channels.

Method	SOTS-Indoor [49]		SOTS-Outdoor [49]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DehazeNet [8]	19.82	0.821	24.75	0.927
AOD-Net [47]	20.51	0.861	24.14	0.920
GridDehazeNet [59]	32.16	0.984	30.86	0.982
MSBDN [25]	33.67	0.985	33.48	0.982
FFA-Net [73]	36.39	0.989	33.57	0.984
ACER-Net [95]	37.17	0.990	-	-
DeHamer [36]	36.63	0.988	35.18	0.986
MAXIM-2S [90]	38.11	0.991	34.19	0.985
PMNet [102]	38.41	0.990	34.74	0.985
DehazeFormer-L [88]	40.05	0.996	-	-
SFNet [19]	41.24	0.996	40.05	0.996
Ours	**41.48**	**0.996**	**40.29**	**0.996**

Table 5. **Image deraining results**. We separately train and evaluate our method on Rain200H, Rain200L, DID-Data, and DDN-Data. PSNR and SSIM scores are calculated on Y channel in YCbCr color space.

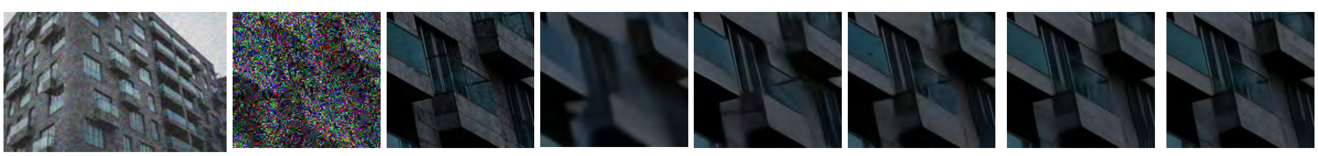
Method	Rain200L [101]		Rain200H [101]		DID-Data [112]		DDN-Data [29]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DDN [28]	34.68	0.967	26.05	0.805	30.97	0.911	30.00	0.904
RESCAN [53]	36.09	0.967	26.75	0.835	33.38	0.941	31.94	0.935
PReNet [78]	37.80	0.981	29.04	0.899	33.17	0.948	32.60	0.946
MSPFN [41]	38.53	0.983	29.36	0.903	33.72	0.955	32.99	0.933
RCDNet [91]	39.17	0.989	30.24	0.904	34.08	0.953	33.04	0.947
MPRNet [110]	39.47	0.982	30.67	0.911	33.99	0.959	33.10	0.935
DualGCN [30]	40.73	0.989	31.15	0.912	34.37	0.962	33.01	0.949
SPDNet [103]	40.50	0.988	31.28	0.920	34.57	0.956	33.15	0.946
Uformer [93]	40.20	0.986	30.80	0.910	35.02	0.962	33.95	0.955
Restormer [111]	40.99	0.989	32.00	0.932	35.29	0.964	34.20	0.957
IDT [98]	40.74	0.988	32.10	**0.934**	34.89	0.962	33.84	0.955
DRSformer [16]	41.21	0.989	**32.16**	0.933	35.24	0.962	34.23	0.955
Ours	**41.59**	**0.990**	31.97	0.931	**35.46**	**0.964**	**34.57**	**0.958**

Table 6. **Grayscale image denoising on Gaussian noise**. Upper-bracket: models are trained on a range of noise levels. Lower-bracket: models are trained on the fixed noise level.

Method	Set12 [115]			BSD68 [63]			Urban100 [39]		
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
DnCNN [115]	32.67	30.35	27.18	31.62	29.16	26.23	32.28	29.80	26.35
FFDNet [117]	32.75	30.43	27.32	31.63	29.19	26.29	32.40	29.90	26.50
IRCNN [116]	32.76	30.37	27.12	31.63	29.15	26.19	32.46	29.80	26.22
DRUNet [119]	33.25	30.94	27.90	31.91	29.48	26.59	33.44	31.11	27.96
Restormer[111]	33.35	31.04	28.01	31.95	29.51	26.62	**33.67**	31.39	28.33
Ours	**33.35**	**31.30**	**28.13**	**31.98**	**29.58**	**26.77**	33.62	**31.47**	**28.46**
FOCNet [40]	33.07	30.73	27.68	31.83	29.38	26.50	33.15	30.64	27.40
MWCNN [58]	33.15	30.79	27.74	31.86	29.41	26.53	33.17	30.66	27.42
NLRN [57]	33.16	30.80	27.64	31.88	29.41	26.47	33.45	30.94	27.49
RNAN [122]	-	-	27.70	-	-	26.48	-	-	27.65
DeamNet [77]	33.19	30.81	27.74	31.91	29.44	26.54	33.37	30.85	27.53
DAGL [65]	33.28	30.93	27.81	31.93	29.46	26.51	33.79	31.39	27.97
SwinIR [55]	33.36	31.01	27.91	**31.97**	29.50	26.58	33.70	31.30	27.98
Restormer [111]	33.42	31.08	28.00	31.96	29.52	26.62	**33.79**	31.46	28.29
Ours	**33.47**	**31.15**	**28.12**	31.92	**29.67**	**26.78**	33.78	**31.58**	**28.38**

Table 7. **Color image denoising on Gaussian noise**. Upper-bracket: models are trained on a range of noise levels. Lower-bracket: models are trained on the fixed noise level. PSNR is calculated on RGB channels.

Method	CBSD68 [64]			Kodak24			McMaster [120]			Urban100 [39]		
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
IRCNN [116]	33.86	31.16	27.86	34.69	32.18	28.93	34.58	32.18	28.91	33.78	31.20	27.70
FFDNet [117]	33.87	31.21	27.96	34.63	32.13	28.98	34.66	32.35	29.18	33.83	31.40	28.05
DnCNN [115]	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
DSNet [70]	33.91	31.28	28.05	34.63	32.16	29.05	34.67	32.40	29.28	-	-	-
DRUNet [119]	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
Restormer [111]	**34.39**	31.78	28.59	35.44	33.02	30.00	35.55	33.31	30.29	**35.06**	32.91	30.02
Ours	34.37	**31.87**	**28.68**	**35.52**	**33.13**	**30.15**	**35.62**	**33.38**	**30.40**	35.03	**32.97**	**30.19**
RPCNN [97]	-	31.24	28.06	-	32.34	29.25	-	32.33	29.33	-	31.81	28.62
BRDNet [89]	34.10	31.43	28.16	34.88	32.41	29.22	35.08	32.75	29.52	34.42	31.99	28.56
RNAN [122]	-	-	28.27	-	-	29.58	-	-	29.72	-	-	29.08
RDN [123]	-	-	28.31	-	-	29.66	-	-	-	-	-	29.38
IPT [12]	-	-	28.39	-	-	29.64	-	-	29.98	-	-	29.71
SwinIR [55]	34.42	31.78	28.56	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82
Restormer [111]	34.40	31.79	28.60	35.47	33.04	30.01	35.61	33.34	30.30	**35.13**	32.96	30.02
Ours	**34.48**	**31.97**	**28.83**	**35.58**	**33.21**	**30.23**	**35.75**	**33.56**	**30.46**	35.11	**33.13**	**30.27**



Input image | Degraded | Label | RDN [123] | RNAN [122] | SwinIR [55] | Restormer [111] | **Ours**

Figure 6. **Color image denoising results on Gaussian noise**. Compared with others, the proposed method can obtain finer results.

Input image | Degraded | Label | DeamNet [77] | RNAN [122] | SwinIR [55] | Restormer [111] | **Ours**

Figure 7. **Grayscale image denoising results on Gaussian noise**. Our method can restore detailed results especially in texture regions.

4.4. Image Deraining Results

Following [16], we train and evaluate the proposed method separately on Rain200L [101], Rain200H [101], DID-Data [112], and DDN-Data [29] datasets. From Table 5, our method achieves +0.31 dB PSNR average improvement on Rain200L, DID-Data, and DDN-Data against DRSformer [16]. On the Rain200H dataset, we only have comparable performance, as its severe rain streak may interfere with the dynamic aggregation process between the guidance

Table 8. **Real-world image denoising results**. We train and evaluate our mehtod on SIDD [1] datasets. * denotes methods using additional training data. PSNR and SSIM scores are calculated on RGB channels.

Dataset	Method	DnCNN [115]	BM3D [20]	CBDNet* [37]	RIDNet* [4]	AINDNet* [44]	VDN [105]	SADNet* [11]	DANet+* [106]	CycleISP* [107]	MIRNet [108]	DeamNet* [77]	MPRNet [110]	DAGL [65]	Uformer [93]	Restormer [111]	**Ours**
SIDD [1]	PSNR ↑	23.66	25.65	30.78	38.71	39.08	39.28	39.46	39.47	39.52	39.72	39.47	39.71	38.94	39.77	<u>40.02</u>	**40.09**
	SSIM ↑	0.583	0.685	0.801	0.951	0.954	0.956	0.957	0.957	0.957	0.959	0.957	0.958	0.953	0.959	<u>0.960</u>	**0.960**

Table 9. Effect of $\mathbf{E}_{txt}$ with different N.

Method	baseline	N=5	N=10	N=20	N=30	N=40
PSNR↑	30.16	31.13	31.36	31.57	31.51	31.60
SSIM↑	0.932	0.941	0.945	0.947	0.947	0.948

Table 10. Effect of integration strategy.

Method	baseline	Enc.	Dec.	Enc. & Dec.
PSNR↑	30.16	31.37	30.31	31.57
SSIM↑	0.932	0.946	0.934	0.947

Table 11. Effect of generated guidance.

Method	baseline	only I2T	Degra.	BLIP2	Ours
PSNR↑	30.16	30.10	30.13	30.34	31.57
SSIM↑	0.932	0.929	0.931	0.940	0.947

information and the degraded images. Visual comparisons of deraining results will be provided in the *Suppl*.

4.5. Image Denoising Results

Image Denoising on Gaussian Noise. Following [111], we train the blind denoising model (σ with range of $[0, 50]$) and non-blind denoising models (σ=$15, 25, 50$) for grayscale and color image denoising on Gaussian noise. We train the proposed method on DFBW dataset with synthetic noise degradation. For grayscale image denoising, we evaluate our method on Set12 [115], BSD68 [64], and Urban100 [39]. For color image denoising, we evaluate our method on CBSD68 [64], Kodak24, McMaster [120], and Urban100 [39]. As shown in Table 6 and Table 7, our method achieves comparable performance with baseline on low noise-level (σ=15), while our method outperforms baseline when noise becomes heavier, *e.g*., achieving +0.25 dB PSNR gain on Urban100 [39] with Gaussian color noise σ=50). Qualitative comparisons in Fig. 6 and Fig. 7 show that our method can restore finer texture and sharper boundaries while other methods tend to generate smoother results.
Real-World Image Denoising. Furthermore, We train and evaluate the proposed method on real-world denoising dataset SIDD [1]. From Table 8, although Restormer [111] has already achieved superior performance on real-world denoising task, our method can also bring +0.07 dB PSNR improvement, demonstrating its effectiveness. Visual comparison results can be found in *Suppl*.

4.6. Ablation Studies

In ablation studies, we adopt NAFNet with 32 channels as our baseline. We train and evaluate each method on GoPro datasets.
Effect of $\mathbf{E}_{txt}$ with different N. In our method, we use the number of words N to control the representation ability of $\mathbf{E}_{txt}$, which is used as condition information to generate guidance. We conduct the experiments on five settings, *i.e*., $N = 5, 10, 20, 30, 40$. As shown in Table 9, the restoration performance generally improves from N=5 to N=40, as more informative textual descriptions can help T2I model generate more faithful guidance. Since the performance only improves slightly when N is larger than 20, we finally set N to 20 to make a trade-off between performance and computational cost.
Effect of Integrating Strategy. To evaluate the effect of different guidance integrating strategies, we conduct experiments by integrating guidance information into different modules of image restoration networks, *i.e*., Enc. (encoder only), Dec. (decoder only), and Enc. & Dec. (both encoder and decoder). As shown in Table 10, the experimental results show that integrating guidance information into both encoder and decoder achieves better performance.
Effect of Proposed Modules. To validate the effectiveness of our proposed modules, we conduct comparison with only using I2T mapper ('only I2T'), using explicit representation ('BLIP2'), and replacing guidance with degraded input ('Degra.'). From comparison results in Table 11, since 'only I2T' and 'Degra.' cannot provide clean information to restoration process, they fail to obtain improvement upon baseline. Since using explicit representation ('BLIP2') cannot keep the content information of input as good as ours, our method can gain more from guidance. Guidance comparisons of 'only I2T', 'BLIP2', and ours can be found in *Suppl*.

5. Conclusion

In this paper, we provide a new perspective for image restoration. Considering that content and degradation are tightly coupled in image representation, while decoupled in textual representation. Rather than direct restoration on image level, we suggest conducting restoration on textual level and in turn utilizing the restored textual content to assist image restoration. To this end, we propose a plug-and-play approach that first encodes degraded images into degraded textual representations and then removes textual degradation information to obtain restored textual representations. Conditioned on the representations, we employ a pre-trained T2I model to synthesize clean guidance images for improving image restoration. We evaluate our method on various image restoration tasks. The experimental results demonstrate it outperforms the state-of-the-art ones.

Acknowledgements. This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. U22B2035 and No. 62371164.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, pages 1692–1700, 2018. 8

[2] Abdullah Abuolaim and Michael S. Brown. Defocus deblurring using dual-pixel data. In *ECCV*, pages 111–126, 2020. 6

[3] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S. Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *ICCV*, pages 2269–2278, 2021. 6

[4] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *ICCV*, pages 3155–3164, 2019. 8

[5] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33 (5):898–916, 2011. 5

[6] Yunpeng Bai, Cairong Wang, Shuzhao Xie, Chao Dong, Chun Yuan, and Zhi Wang. Textir: A simple framework for text-based editable image restoration. *arXiv preprint arXiv:2302.14736*, 2023. 3

[7] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions, 2023. https://cdn.openai.com/papers/dall-e-3. 3

[8] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Trans. Image Process.*, 25 (11):5187–5198, 2016. 1, 7

[9] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *ECCV*, pages 325–342, 2022. 3

[10] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3

[11] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. In *ECCV*, pages 171–187, 2020. 8

[12] 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, 6, 7

[13] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *CVPRW*, pages 182–192, 2021. 1, 6

[14] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, pages 17–33, 2022. 1, 5, 6

[15] 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 17632–17641, 2022. 1, 3

[16] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning A sparse transformer network for effective image deraining. In *CVPR*, pages 5896–5905, 2023. 1, 5, 7

[17] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023. 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 *ICCV*, pages 4621–4630, 2021. 1, 6

[19] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *ICLR*, 2023. 1, 5, 7

[20] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8):2080–2095, 2007. 8

[21] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, 34:19822–19835, 2021. 3

[22] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 3

[23] Yuxuan Ding, Chunna Tian, Haoxuan Ding, and Lingqiao Liu. The clip model is secretly an image-to-prompt converter. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[24] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014. 2

[25] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2154–2164, 2020. 7

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[27] Qingnan Fan, Dongdong Chen, Lu Yuan, Gang Hua, Nenghai Yu, and Baoquan Chen. A general decoupled learning framework for parameterized image operators. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):33–47, 2021. 5

[28] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John W. Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 1715–1723, 2017. 7

[29] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John W. Paisley. Removing rain from single

images via a deep detail network. In *CVPR*, pages 1715–1723, 2017. 7

[30] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. Rain streak removal via dual graph convolutional network. In *AAAI*, pages 1352–1360, 2021. 7

[31] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, pages 89–106, 2022. 3

[32] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, pages 3848–3856, 2019. 5

[33] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, pages 3848–3856, 2019. 5

[34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014. 3

[35] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *ICCV*, pages 2511–2520, 2019. 2

[36] Chunle Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, pages 5802–5810, 2022. 1, 7

[37] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, pages 1712–1722, 2019. 8

[38] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 3

[39] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 7, 8

[40] Xixi Jia, Sanyang Liu, Xiangchu Feng, and Lei Zhang. Focnet: A fractional optimal control network for image denoising. In *CVPR*, pages 6054–6063, 2019. 7

[41] 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 *CVPR*, pages 8343–8352, 2020. 7

[42] Yuming Jiang, Kelvin C. K. Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, pages 2103–2112, 2021. 3

[43] Ali Karaali and Cláudio R. Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Trans. Image Process.*, 27(3):1126–1137, 2018. 6

[44] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *CVPR*, pages 3479–3489, 2020. 8

[45] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *CVPR*, pages 12222–12230, 2019. 6

[46] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, pages 2034–2042, 2021. 6

[47] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, pages 4780–4788, 2017. 7

[48] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *NeurIPS*, 32, 2019. 3

[49] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.*, 28(1):492–505, 2019. 5, 6, 7

[50] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17431–17441, 2022. 1, 3, 5

[51] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 1

[52] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 1, 4

[53] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, pages 262–277, 2018. 7

[54] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, pages 262–277, 2018. 2

[55] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Int. Conf. Comput. Vis. Worksh.*, pages 1833–1844, 2021. 1, 3, 7

[56] 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

[57] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S. Huang. Non-local recurrent network for image restoration. In *NeurIPS*, pages 1680–1689, 2018. 2, 7

[58] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *CVPRW*, pages 773–782, 2018. 7

[59] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, pages 7313–7322. IEEE, 2019. 7

[60] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[61] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. MASA-SR: matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, pages 6368–6377, 2021. 3, 5

[62] 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 Trans. Image Process.*, 26(2): 1004–1016, 2017. 5

[63] David R. Martin, Charless C. 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 416–425, 2001. 7

[64] David R. Martin, Charless C. 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 416–425, 2001. 7, 8

[65] Chong Mou, Jian Zhang, and Zhuoyuan Wu. Dynamic attentive graph learning for image restoration. In *ICCV*, pages 4308–4317, 2021. 7, 8

[66] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 257–265, 2017. 6

[67] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[68] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*, pages 327–343, 2020. 6

[69] 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. 1, 3

[70] Yali Peng, Lu Zhang, Shigang Liu, Xiaojun Wu, Yu Zhang, and Xili Wang. Dilated residual networks with symmetric skip connection for image denoising. *Neurocomputing*, 345:67–76, 2019. 7

[71] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. In *NeurIPS*, 2023. 1, 2, 3, 5

[72] Kuldeep Purohit, Maitreya Suin, A. N. Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *ICCV*, pages 2289–2299, 2021. 6

[73] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, pages 11908–11915, 2020. 7

[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 4

[75] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[76] 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

[77] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *CVPR*, pages 8596–8606, 2021. 7, 8

[78] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, pages 3937–3946, 2019. 7

[79] 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. 6

[80] 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. 3

[81] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 2, 4

[82] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[83] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 3

[84] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, pages 5571–5580, 2019. 6

[85] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *CVPR*, pages 657–665, 2015. 6

[86] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *ICCV*, pages 2622–2630, 2021. 6

[87] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[88] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Trans. Image Process.*, 32:1927–1941, 2023. 1, 7

[89] Chunwei Tian, Yong Xu, and Wangmeng Zuo. Image denoising using deep CNN with batch renormalization. *Neural Networks*, 121:461–473, 2020. 5, 7

[90] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan C. Bovik, and Yinxiao Li. MAXIM: multi-axis MLP for image processing. In *CVPR*, pages 5759–5770, 2022. 1, 7

[91] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *CVPR*, pages 3100–3109, 2020. 7

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

[93] 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 17662–17672, 2022. 1, 2, 6, 7, 8

[94] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 4

[95] 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. 7

[96] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 3

[97] Zhihao Xia and Ayan Chakrabarti. Identifying recurring patterns with deep neural networks for natural image denoising. In *WACV*, pages 2415–2423, 2020. 7

[98] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):12978–12995, 2023. 1, 7

[99] Liangbin Xie, Xintao Wang, Chao Dong, Zhongang Qi, and Ying Shan. Finding discriminative filters for specific degradations in blind super-resolution. In *NeurIPS*, pages 51–61, 2021. 3

[100] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 3

[101] Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, pages 1685–1694, 2017. 7

[102] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *ECCV*, pages 130–145. Springer, 2022. 7

[103] Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, and Tieyong Zeng. Structure-preserving deraining with residue channel prior guidance. In *ICCV*, pages 4218–4227, 2021. 7

[104] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[105] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational denoising network: Toward blind noise modeling and removal. In *NeurIPS*, pages 1688–1699, 2019. 8

[106] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *ECCV*, pages 41–58, 2020. 8

[107] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *CVPR*, pages 2693–2702, 2020. 8

[108] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, pages 492–511, 2020. 8

[109] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, pages 492–511, 2020. 2

[110] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 1, 2, 5, 6, 7, 8

[111] 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 5718–5729, 2022. 1, 3, 5, 6, 7, 8

[112] He Zhang and Vishal M. Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, pages 695–704, 2018. 7

[113] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, pages 5978–5986, 2019. 6

[114] 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. 1, 3

[115] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.*, 26(7):3142–3155, 2017. 1, 2, 7, 8

[116] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, pages 2808–2817, 2017. 7

[117] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Trans. Image Process.*, 27(9):4608–4622, 2018. 1, 7

[118] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Björn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, pages 2734–2743, 2020. 6

[119] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6360–6376, 2022. 1, 7

[120] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and non-local adaptive thresholding. *J. Electronic Imaging*, 20(2): 023016, 2011. 7, 8

[121] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 294–310, 2018. 2

[122] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 2, 7

[123] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(7):2480–2495, 2021. 7

[124] 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. 1, 3