# High-Quality Image Restoration Following Human Instructions

Marcos V. Conde [1,2], Gregor Geigle [1], Radu Timofte [1]
[1] Computer Vision Lab, CAIDAS & IFI, University of Würzburg
[2] Sony PlayStation, FTG
https://github.com/mv-lab/InstructIR


Figure 1. Given an **image** and a **prompt** for how to improve that image, our *all-in-one* restoration model corrects the image considering the human instruction. *InstructIR*, can tackle various types and levels of degradation, and it is able to generalize to restoring *real* photos.

## Abstract

*Image restoration is a fundamental problem that involves recovering a high-quality clean image from its degraded observation. All-In-One image restoration models can effectively restore images from various types and levels of degradation using degradation-specific information as prompts to guide the restoration model. In this work, we present the first approach that uses human-written instructions to guide the image restoration model. Given natural language prompts, our model can recover high-quality images from their degraded counterparts, considering multiple degradation types. Our method, InstructIR, achieves state-of-the-art results on several restoration tasks including image denoising, deraining, deblurring, dehazing, and (low-light) image enhancement. InstructIR improves +1dB over previous all-in-one restoration methods. Moreover, our dataset and results represent a novel benchmark for new research on text-guided image restoration and enhancement.*

## 1. Introduction

Images often contain unpleasant effects such as noise, motion blur, haze, and low dynamic range. Such effects are commonly known in low-level computer vision as *degradations*. These can result from camera limitations or challenging environmental conditions *e.g*. low light.

Image restoration aims to recover a high-quality image from its degraded counterpart. This is a complex inverse problem since multiple different solutions can exist for restoring any given image [16, 20, 44, 59, 102, 103].

Some methods focus on specific degradations, for instance reducing noise (denoising) [64, 102, 103], removing blur (deblurring) [58, 105], or clearing haze (dehazing) [16, 66]. Such methods are effective for their specific task, yet they do not generalize well to other types of degradation. Other approaches use a general neural network for diverse tasks [10, 74, 82, 95], yet training the neural network for each specific task independently. Since using a separate model for each possible degradation is resourceintensive, recent approaches propose *All-in-One* restoration models [42, 60, 61, 100]. These approaches use a single deep blind restoration model considering multiple degradation types and levels. Contemporary works such as PromptIR [61] or ProRes [49] utilize a unified model for blind image restoration using learned guidance vectors, also known as "prompt *embeddings*", in contrast to raw user prompts in text form, which we use in this work.

In parallel, recent works such as InstructPix2Pix [4] show the potential of using text prompts to guide image generation and editing models. However, this method (or recent alternatives) do not tackle inverse problems. Inspired by these works, we argue that text guidance can help to guide blind restoration models better than the image-based degradation classification used in previous works [42, 60, 100]. Users generally have an idea about what has to be fixed (though they might lack domain-specific vocabulary) so we can use this information to guide the model.

**Contributions** We propose the first approach that utilizes real human-written instructions to solve inverse problems and image restoration. Our comprehensive experiments demonstrate the potential of using text guidance for image restoration and enhancement by achieving *state-of-the-art* performance on various image restoration tasks, including image denoising, deraining, deblurring, dehazing and low-light image enhancement. Our model, *InstructIR*, is able to generalize to restoring images using arbitrary human-written instructions. Moreover, our single *all-in-one* model covers more tasks than many previous works. We show diverse restoration samples of our method in Figure 1.

## 2. Related Work

**Image Restoration.** Recent deep learning methods [16, 44, 58, 64, 74, 95] have shown consistently better results compared to traditional techniques for blind image restoration [18, 29, 35, 37, 54, 73]. The proposed neural networks are based on convolutional neural networks (CNNs) and Transformers [76] (or related attention mechanisms).

We focus on general-purpose restoration models [10, 44, 82, 95]. For example, SwinIR [44], MAXIM [74] and Uformer [82]. These models can be trained -independently- for diverse tasks such as denoising, deraining or deblurring. Their ability to capture local and global feature interactions, and enhance them, allows the models to achieve great performance consistently across different tasks. For instance, Restormer [95] uses non-local blocks [79] to capture complex features across the image.

NAFNet [10] is an efficient alternative to complex transformer-based methods. The model uses simplified channel attention, and gating as an alternative to non-linear activations. The builing block (NAFBlock) follows a simple meta-former [92] architecture with efficient inverted residual blocks [31]. In this work, we build our *InstructIR* model using NAFNet as backbone, due to its efficient and simple design, and high performance in several restoration tasks.

**All-in-One Image Restoration.** Single degradation (or single task) restoration methods are well-studied, however, their real-world applications are limited due to the required resources *i.e.* allocating different models, and select the adequate model on demand. Moreover, images rarely present a single degradation, for instance noise and blur are almost ubiquitous in any image capture.

All-in-One (also known as multi-degradation or multi-task) image restoration is emerging as a new research field in low-level computer vision [42, 49, 60, 61, 75, 91, 97, 98]. These approaches use a single deep blind restoration model to tackle different degradation types and levels.

We use as reference AirNet [42], IDR [100] and ADMS [60]. We also consider the contemporary work PromptIR [61]. The methods use different techniques to guide the blind model in the restoration process. For instance, an auxiliary model for degradation classification [42, 60], or multi-dimensional guidance vectors (also known as "prompts") [49, 61] that help the model to discriminate the different types of degradation in the image.

Despite it is not the focus of this work, we acknowledge that *real-world image super-resolution* is a related problem [12, 44, 48, 106], since the models aim to solve an inverse problem considering multiple degradations (blur, noise and downsampling).

**Text-guided Image Manipulation.** In the recent years, multiple methods have been proposed for text-to-image generation and text-based image editing works [4, 30, 34, 53, 70]. These models use text prompts to describe images or actions, and powerful diffusion-based models for generating the corresponding images. Our main reference is InstructPix2Pix [4], this method enables editing from *instructions* that tell the model what action to perform, as opposed to text labels, captions or descriptions of the input or output images. Therefore, the user can transmit what to do in natural written text, without requiring to provide further image descriptions or sample reference images.

## 3. Image Restoration Following Instructions

We treat instruction-based image restoration as a supervised learning problem similar to previous works [4]. First, we generate over 10000 prompts using GPT-4 based on our own sample instructions. We explain the creation of the prompt dataset in Sec. 3.1. We then build a large paired training dataset of prompts and degraded/clean images. Finally, we train our *InstructIR* model, and we evaluate it on a wide variety of instructions including real human-written prompts. We explain our text encoder in Sec 3.2, and our complete model in Sec. 3.3.

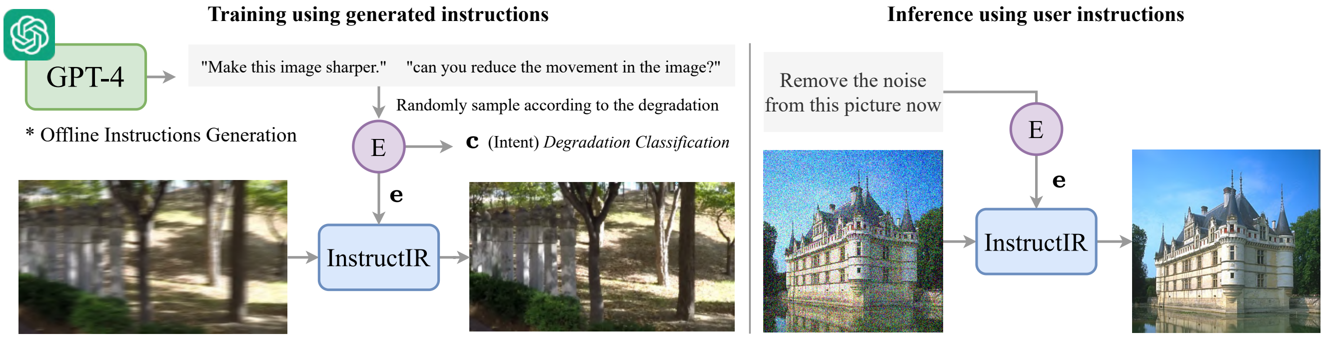



Figure 2. We train our blind image restoration models using common image datasets, and prompts generated using GPT-4, note that this is (self-)supervised learning. At inference time, our model generalizes to human-written instructions and restores real images.

## 3.1. Generating Prompts for Training

**Why instructions?** Inspired by InstructPix2Pix [4], we adopt human written instructions as the mechanism of control for our model. There is no need for the user to provide additional information, such as example clean images, or descriptions of the visual content. Instructions offer a clear and expressive way to interact, enabling users to pinpoint the unpleasant effects (degradations) in the images.

Handling free-form user prompts rather than fixed degradation-specific prompts increases the usability of our model for laypeople who lack domain expertise. We thus want our model to be capable of understanding diverse prompts posed by users "in-the-wild" *e.g.* kids, adults, or photographers. To this end, we use a large language model (*i.e.*, GPT-4) to create diverse requests that might be asked by users for the different degradations types. We then filter those generated prompts to remove ambiguous or unclear prompts (*e.g.*, *"Make the image cleaner", "improve this image"*). Our final instructions set contains over 10000 different prompts in total, for 7 different tasks. We display some examples in Table 1. As we show in Figure 2 the prompts are sampled randomly depending on the input degradation.

## 3.2. Text Encoder

**The Choice of the Text Encoder.** A text encoder maps the user prompt to a fixed-size vector representation (a text embedding). The related methods for text-based image generation [67] and manipulation [3, 4] often use the text encoder of a CLIP model [62] to encode user prompts as CLIP excels in visual prompts. However, user prompts for degradation contain, in general, little to no visual content (*e.g.* the use describes the degradation, not the image itself), therefore, the large CLIP encoders (with over 60 million parameters) are not suitable – especially if we require efficiency.

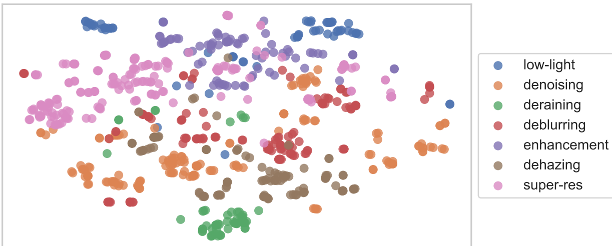
We opt, instead, to use a pure text-based sentence encoder [63], that is, a model trained to encode sentences in a semantically meaningful embedding space. Sentence encoders –pre-trained with millions of examples– are compact and fast in comparison to CLIP, while being able to encode the semantics of diverse user prompts. For instance, we use the `BGE-micro-v2` sentence transformer.

Table 1. Examples of our curated GPT4-generated user prompts with varying language and domain expertise.

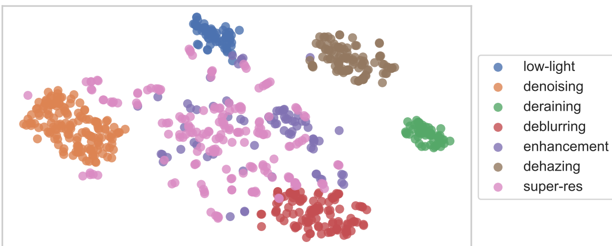
| Degradation | Prompts |
|---|---|
| Denoising | Can you clean the dots from my image?<br>Fix the grainy parts of this photo<br>Remove the noise from my picture |
| Deblurring | Can you reduce the movement in the image?<br>My picture's not sharp, fix it<br>Deblur my picture, it's too fuzzy |
| Dehazing | Can you make this picture clearer?<br>Help, my picture is all cloudy<br>Remove the fog from my photo |
| Deraining | I want my photo to be clear, not rainy<br>Clear the rain from my picture<br>Remove the raindrops from my photo |
| Super-Res. | Make my photo bigger and better<br>Add details to this image<br>Increase the resolution of this photo |
| Low-light | The photo is too dark, improve exposure<br>Increase the illumination in this shot<br>My shot has very low dynamic range |
| Enhancement | Make it pop!<br>Adjust the color balance for a natural look<br>Apply a cinematic color grade to the photo |

**Fine-tuning the Text Encoder.** We want to adapt the text encoder $\mathrm{E}$ for the restoration task to better encode the required information for the restoration model. Training the full text encoder is likely to lead to overfitting on our small training set and lead to loss of generalization. Instead, we freeze the text encoder and train a projection head on top:

$$\mathbf{e} = \mathrm{norm}(\mathbf{W} \cdot \mathrm{E}(t)) \quad (1)$$




(a) t-SNE of embeddings *before* training *i.e.* frozen text encoder




(b) t-SNE of embeddings *after* training our learned projection

Figure 3. We show t-SNE plots of the text embeddings before/after training *InstructIR*. Each dot represents a human instruction.

where $t$ is the text, $\mathrm{E}(t)$ represents the raw text embedding, $\mathbf{W} \in \mathbb{R}^{d_t \times d_v}$ is a learned projection from the text dimension ($d_t$) to the input dimension for the restoration model ($d_v$), and $\mathrm{norm}$ is the l2-norm.

Figure 3 shows that while the text encoder is capable out-of-the-box to cluster instructions to some extent (Figure 3a), our trained projection yields greatly improved clusters (Figure 3b). We distinguish clearly the clusters for deraining, denoising, dehazing, deblurring, and low-light image enhancement. The instructions for such tasks or degradations are very characteristic. Furthermore, we can appreciate that "super-res" and "enhancement" tasks are quite spread and between the previous ones, which matches the language logic. For instance *"add details to this image"* could be used for enhancement, deblurring or denosising. In our experiments, $d_t = 384$, $d_v = 256$ and $\mathbf{W}$ is a linear layer.

**Intent Classification Loss.** We propose a guidance loss on the text embedding $\mathbf{e}$ to improve training and interpretability. Using the degradation types as targets, we train a simple classification head $\mathcal{C}$ such that $\mathbf{c} = \mathcal{C}(\mathbf{e})$, where $\mathbf{c} \in \mathrm{R}^D$, being $D$ is the number of degradations. In our experiments $D = 7$ and the classification head $\mathcal{C}$ is a simple two-layers MLP. Thus, we only need to train a projection layer and a simple MLP to capture the natural language knowledge. This allows the text model to learn meaningful embeddings as we can appreciate in Figure 3, not just guidance vectors for the main image processing model.

We find that the model is able to classify accurately (*i.e.* over 95% accuracy) the underlying degradation in the user's prompt after a few epochs.

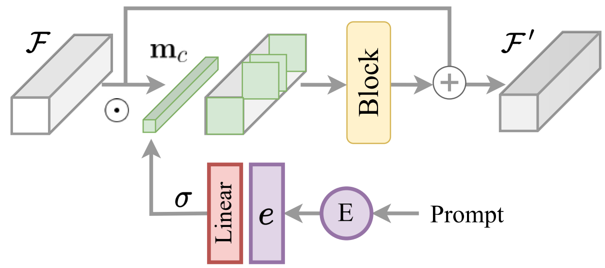



Figure 4. *Instruction Condition Block (ICB)* using an approximation of task routing [71] for many-tasks learning. See Eq. 2.

### 3.3. InstructIR

Our method *InstructIR* consists of an image model and a text encoder. We introduced our text encoder in Sec. 3.2. We use NAFNet [10] as the image model, an efficient image restoration model that follows a U-Net architecture [68]. To successfully learn multiple tasks using a single model, we use task routing techniques. Our framework for training and evaluating the model is illustrated in Figure 2.

**Text Guidance.** The key aspect of *InstructIR* is the integration of the encoded instruction as a mechanism of control for the image model. Inspired in *task routing* for many-task learning [14, 69, 71], we propose an *"Instruction Condition Block" (ICB)* to enable task-specific transformations within the model. Conventional task routing [71] applies task-specific binary masks to the channel features. Since our model does not know *a-priori* the degradation, we cannot use this technique directly.

Considering the image features $\mathcal{F}$, and the encoded instruction $\mathbf{e}$, we apply task routing as follows:

$$\mathcal{F}'_c = \mathrm{Block}(\mathcal{F}_c \odot \mathbf{m}_c) + \mathcal{F}_c \tag{2}$$

where the mask $\mathbf{m}_c = \sigma(\mathbf{W_c} \cdot \mathbf{e})$ is produced using a linear layer -activated using the Sigmoid function- to produce a set of weights depending on the text embedding $\mathbf{e}$. Thus, we obtain a $c$-dimensional per-channel (soft-)binary mask $\mathbf{m}_c$. As [71], task routing is applied as the channel-wise multiplication $\odot$ for masking features depending on the task. The conditioned features are further enhanced using a NAFBlock [10] (Block). We illustrate our task-routing ICB block in Figure 4. We use "regular" NAFBlocks [10], followed by ICBs to condition the features, at both encoder and decoder blocks. The formulation is $F^{l+1} = \mathrm{ICB}(\mathrm{Block}(F^l))$ where $l$ is the layer. Although we do not condition explicitly the filters of the neural network, as in [71], the mask allows the model to select the most relevant channels depending on the image information and the instruction. Note that this formulation enables differentiable feature masking, and certain interpretability *i.e.* the features with high weights contribute the most to the restoration process. Indirectly, this also enforces to learn diverse filters and reduce sparsity [14, 71].

**Is *InstructIR* a blind restoration model?** The model does not use explicit information about the degradation in the image *e.g.* noise profiles, blur kernels, or PSFs. Since our model infers the task (degradation) given the image and the instruction, we consider *InstructIR* a *blind* image restoration model. Similarly to previous works that use auxiliary image-based degradation classification [42, 60].

# 4. Experimental Results

We provide extensive qualitative results using benchmark images in Figures 17, 18 and 19. We also evaluate our model on 9 well-known benchmarks for different image restoration tasks: image denoising, deblurring, deraining, dehazing, and image enhancement. We present extensive quantitative results in Table 2. Our *single* model successfully restores images considering different degradation types and levels. We provide additional results and ablation studies in the supplementary material.

## 4.1. Implementation Details.

Our *InstructIR* model is end-to-end trainable. The image model does not require pre-training, yet we use a pre-trained sentence encoder as language model.

**Text Encoder.** As we discussed in Sec. 3.2, we only need to train the text embedding projection and classification head ($\approx 100K$ parameters). We initialize the text encoder with BGE-MICRO-V2 [1], a distilled version of BGE-SMALL-EN [85]. The BGE encoders are BERT-like encoders [13] pre-trained on large amounts of supervised and unsupervised data for general-purpose sentence encoding. The BGE-micro model is a 3-layer encoder with 17.3 million parameters, which we freeze during training. We also explore ALL-MINILM-L6-V2 and CLIP encoders, however, we concluded that small models prevent overfitting and provide the best performance while being fast. We provide the ablation study comparing the three text encoders in the supplementary material.

**Image Model.** We use NAFNet [10] as image model. The architecture consists of a 4-level encoder-decoder, with varying numbers of blocks at each level, specifically [2, 2, 4, 8] for the encoder, and [2, 2, 2, 2] for the decoder, from the level-1 to level-4 respectively. Between the encoder and decoder we use 4 middle blocks to enhance further the features. The decoder implements addition instead of concatenation for the skip connections.

We use the *Instruction Condition Block (ICB)* for task-routing [71] only in the encoder and decoder.

The model is optimized using the $\mathcal{L}_1$ loss between the ground-truth clean image and the restored one. Additionally we use the cross-entropy loss $\mathcal{L}_{ce}$ for the intent classification head of the text encoder. We train use a batch size of 32 and AdamW [36] optimizer with learning rate $5e^{-4}$ for 500 epochs (approximately 1 day using a single NVIDIA A100). We also use cosine annealing learning rate decay. During training, we utilize cropped patches of size $256 \times 256$ as input, and we use random horizontal and vertical flips as augmentations. Since our model uses as input instruction-image pairs, given an image, and knowing its degradation, we randomly sample instructions from our prompt dataset ($>10$K samples). Our image model has only 16M parameters, and the learned text projection is just 100k parameters (the language model is 17M parameters), thus, our model can be trained easily on standard GPUs such as NVIDIA RTX 2080Ti or 3090Ti in a couple of days. Furthermore, the inference process also fits in low-computation budgets.

[1] https://huggingface.co/TaylorAI/bge-micro-v2

## 4.2. Datasets and Benchmarks

Following previous works [42, 61, 100], we prepare the datasets for different restoration tasks.

**Image denoising.** We use a combination of BSD400 [2] and WED [50] datasets for training. This combined training set contains $\approx 5000$ images. Using as reference the clean images in the dataset, we generate the noisy images by adding Gaussian noise with different noise levels $\sigma \in \{15, 25, 50\}$. We test the models on the well-known BSD68 [52] and Urban100 [32] datasets.

**Image deraining.** We use the Rain100L [88] dataset, which consists of 200 clean-rainy image pairs for training, and 100 pairs for testing.

**Image dehazing.** We utilize the Reside (outdoor) SOTS [41] dataset, which contains $\approx$ 72K training images. However, many images are low-quality and unrealistic, thus, we filtered the dataset and selected a random set of 2000 images – also to avoid imbalance *w.r.t* the other tasks. We use the standard *outdoor* testset of 500 images.

**Image deblurring.** We use the GoPro dataset for motion deblurring [57] which consist of 2103 images for training, and 1111 for testing.

**Low-light Image Enhancement.** We use the LOL [83] dataset (v1), and we adopt its official split of 485 training images, and 15 testing images.

**Image Enhancement.** Extending previous works, we also study photo-realistic image enhancement using the MIT5K dataset [5]. We use 1000 images for training, and the standard split of 500 images for testing (as in [74]).

Finally, as previous works [42, 61, 100], we combine all the aforementioned training datasets, and we train our unified model for all-in-one restoration.

Table 2. Quantitative results on **five restoration tasks (5D)** with *state-of-the-arts* general image restoration and all-in-one methods. We highlight the reference model *without* text, the best results, and the second best results. We also present the ablation study of our *multi-task variants* (from 5 to 7 tasks — 5D, 6D, 7D). This table is based on Zhang *et al.* IDR [100] (*CVPR '23*).

| Methods | Deraining Rain100L [88] | | Dehazing SOTS [41] | | Denoising BSD68 [52] | | Deblurring GoPro [57] | | Low-light Enh. LOL [83] | | Average | | Params |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | (M) |
| HINet [9] | 35.67 | 0.969 | 24.74 | 0.937 | 31.00 | 0.881 | 26.12 | 0.788 | 19.47 | 0.800 | 27.40 | 0.875 | 88.67 |
| DGUNet [56] | 36.62 | 0.971 | 24.78 | 0.940 | 31.10 | 0.883 | 27.25 | 0.837 | 21.87 | 0.823 | 28.32 | 0.891 | 17.33 |
| MIRNetV2 [93] | 33.89 | 0.954 | 24.03 | 0.927 | 30.97 | 0.881 | 26.30 | 0.799 | 21.52 | 0.815 | 27.34 | 0.875 | 5.86 |
| SwinIR [44] | 30.78 | 0.923 | 21.50 | 0.891 | 30.59 | 0.868 | 24.52 | 0.773 | 17.81 | 0.723 | 25.04 | 0.835 | 0.91 |
| Restormer [95] | 34.81 | 0.962 | 24.09 | 0.927 | 31.49 | 0.884 | 27.22 | 0.829 | 20.41 | 0.806 | 27.60 | 0.881 | 26.13 |
| NAFNet [10] | 35.56 | 0.967 | 25.23 | 0.939 | 31.02 | 0.883 | 26.53 | 0.808 | 20.49 | 0.809 | 27.76 | 0.881 | 17.11 |
| DL [21] | 21.96 | 0.762 | 20.54 | 0.826 | 23.09 | 0.745 | 19.86 | 0.672 | 19.83 | 0.712 | 21.05 | 0.743 | 2.09 |
| Transweather [75] | 29.43 | 0.905 | 21.32 | 0.885 | 29.00 | 0.841 | 25.12 | 0.757 | 21.21 | 0.792 | 25.22 | 0.836 | 37.93 |
| TAPE [45] | 29.67 | 0.904 | 22.16 | 0.861 | 30.18 | 0.855 | 24.47 | 0.763 | 18.97 | 0.621 | 25.09 | 0.801 | 1.07 |
| AirNet [42] | 32.98 | 0.951 | 21.04 | 0.884 | 30.91 | 0.882 | 24.35 | 0.781 | 18.18 | 0.735 | 25.49 | 0.846 | 8.93 |
| IDR [100] | 35.63 | 0.965 | 25.24 | 0.943 | 31.60 | 0.887 | 27.87 | 0.846 | 21.34 | 0.826 | <u>28.34</u> | <u>0.893</u> | 15.34 |
| *InstructIR*-5D | 36.84 | 0.973 | 27.10 | 0.956 | 31.40 | 0.887 | 29.40 | 0.886 | 23.00 | 0.836 | **29.55** | **0.907** | 15.8 |
| *InstructIR*-6D | 36.80 | 0.973 | 27.00 | 0.951 | 31.39 | 0.888 | 29.73 | 0.892 | 22.83 | 0.836 | **29.55** | **0.908** | 15.8 |
| *InstructIR*-7D | 36.75 | 0.972 | 26.90 | 0.952 | 31.37 | 0.887 | 29.70 | 0.892 | 22.81 | 0.836 | 29.50 | 0.907 | 15.8 |

### 4.3. Multiple Degradation Results

We define two initial setups for multi-task restoration:

- **3D** for *three-degradation* models such as AirNet [42], these tackle image denoising, dehazing and deraining.
- **5D** for *five-degradation* models, considering image denoising, deblurring, dehazing, deraining and low-light image enhancement as in [100].

In Table 2, we show the performance of **5D** models. Following Zhang *et al.* [100], we compare *InstructIR* with several *state-of-the-art* methods for general image restoration [9, 10, 44, 93, 95], and all-in-one image restoration methods [21, 42, 45, 75, 100]. We can observe that our simple image model (just 16M parameters) can tackle successfully at least five different tasks thanks to the instruction-based guidance, and achieves the most competitive results.

In Table 4 we can appreciate a similar behaviour, when the number of tasks is just three (**3D**), our model improves further in terms of reconstruction performance.

Based on these results, we pose the following question: *How many tasks can we tackle using a single model without losing too much performance?* To answer this, we propose the **6D** and **7D** variants. For the **6D** variant, we fine-tune the original **5D** to consider also super-resolution as sixth task. Finally, our **7D** model includes all previous tasks, and additionally image enhancement (MIT5K photo retouching). We show the performance of these two variants in Table 2.

**Test Instructions.** *InstructIR* requires as input the degraded image <u>and</u> the human-written instruction. Therefore, we also prepare a testset of prompts *i.e.* instruction-image test pairs. The performance of *InstructIR* depends on the ambiguity and precision of the instruction. We provide the ablation study in Table 3. *InstructIR* is quite robust to more/less detailed instructions. However, it is still limited with ambiguous instructions such as *"enhance this image"*. We show diverse instructions in the following Figures.

Table 3. Ablation study on the *sensitivity of instructions*. We report PSNR/SSIM metrics for each task using our **5D** base model. We repeat the evaluation on each testset 10 times, each time we sample different prompts for each image, and we report the average results. The "Real Users †" in this study are amateur photographers, thus, the instructions were very precise.

| Language Level | Deraining | Denoising | Deblurring | LOL |
|---|---|---|---|---|
| Basic & Precise | 36.84/0.973 | 31.40/0.887 | 29.47/0.887 | 23.00/0.836 |
| Basic & Ambiguous | 36.24/0.970 | 31.35/0.887 | 29.21/0.885 | 21.85/0.827 |
| Real Users † | 36.84/0.973 | 31.40/0.887 | 29.47/0.887 | 23.00/0.836 |

## 5. Multi-Task Ablation Study

***How does 6D work?*** Besides the 5 basic tasks -as previous works-, we include single image super-resolution (SISR). For this, we include as training data the DIV2K [1]. Since our model does not perform upsampling, we use the Bicubic degradation model [1, 15] for generating the low-resolution images (LR), and the upsampled versions (HR) that are fed into our model to enhance them. Adding this extra task increases the performance on deblurring –a related degradation–, without harming notably the performance on the other tasks. However, the performance on SR benchmarks is far from classical super-resolution methods [1, 44].

***How does 7D work?*** Finally, if we add image enhancement –a task not related to the previous ones *i.e.* inverse problems– the performance on the restoration tasks decays slightly. However, the model still achieves *state-of-the-art* results. Moreover, as we show in Table 5, the performance

Table 4. Comparisons of all-in-one restoration models for **three restoration tasks (3D)**. We also show an ablation study for image denoising -the fundamental inverse problem- considering different noise levels. We report PSNR/SSIM metrics. Table based on [61].

| Methods | Dehazing SOTS [41] | Deraining Rain100L [21] | Denoising ablation study (BSD68 [52]) $\sigma = 15$ | $\sigma = 25$ | $\sigma = 50$ | Average |
|---|---|---|---|---|---|---|
| BRDNet [72] | 23.23/0.895 | 27.42/0.895 | 32.26/0.898 | 29.76/0.836 | 26.34/0.836 | 27.80/0.843 |
| LPNet [25] | 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 [19] | 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 [94] | 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[21] | 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 [42] | 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 [61] | **30.58/0.974** | 36.37/0.972 | 33.98/0.933 | 31.31/0.888 | 28.06/0.799 | 32.06/0.913 |
| *InstructIR*-**3D** | 30.22/0.959 | **37.98/0.978** | **34.15/0.933** | **31.52/0.890** | **28.30/0.804** | **32.43/0.913** |
| *InstructIR*-**5D** | 27.10/0.956 | 36.84/0.973 | 34.00/0.931 | 31.40/0.887 | 28.15/0.798 | 31.50/0.909 |

Rain, Blur and Noise *"Correct the noise"* *"Remove the rain"* *"Increase resolution"* *"Enhance the photo"*

Figure 5. **Selective task**. *InstructIR* can remove particular degradations or perform different transformations depending on the human instructions. This is a novel feature in image restoration, and it is possible thanks to the novel integration of textual descriptions.

Table 5. **Image Enhancement** performance on MIT5K [5, 96].

| Method | PSNR ↑ | SSIM ↑ | $\Delta E_{ab}$ ↓ |
|---|---|---|---|
| UPE [77] | 21.88 | 0.853 | 10.80 |
| DPE [26] | 23.75 | 0.908 | 9.34 |
| HDRNet [11] | 24.32 | 0.912 | 8.49 |
| 3DLUT [96] | **25.21** | **0.922** | **7.61** |
| *InstructIR*-**7D** | 24.65 | 0.900 | 8.20 |

Table 6. **Summary ablation study** on the multi-task variants of *InstructIR* that tackle from 3 to 7 tasks. We report PSNR/SSIM.

| Tasks | Rain | Noise ($\sigma 15$) | Blur | LOL |
|---|---|---|---|---|
| **3D** | 37.98/0.978 | 31.52/0.890 | - | - |
| **5D** | 36.84/0.973 | 31.40/0.887 | 29.40/0.886 | 23.00/0.836 |
| **6D** | 36.80 0.973 | 31.39 0.888 | 29.73/0.892 | 22.83 0.836 |
| **7D** | 36.75 0.972 | 31.37 0.887 | 29.70/0.892 | 22.81 0.836 |

on this task using the MIT5K [5] Dataset is notable, while keeping the performance on the other tasks. We achieve similar performance to classical task-specific methods.

We **summarize** the multi-task ablation study in Table 6. Our model can tackle multiple tasks without losing performance notably thanks to the instruction-based task routing.

**Comparison with Task-specific Methods** Our main goal is to design a powerful all-in-one model, thus, *InstructIR* was not designed to be trained for a particular degradation. Nevertheless, we also compare *InstructIR* with task-specific methods *i.e*. models tailored and trained for specific tasks.

We compare with task-specific methods for image enhancement in Table 5, and for low-light in image enhancement in 7. We provide extensive comparisons for image denoising in Table 8. Also, in Table 9 we show comparisons with classical methods for deblurring and dehazing. Our multi-task method is better than most task-specific methods, yet it is still not better than SOTA.

## 6. On the Effectiveness of Instructions

Thanks to our integration of human instructions, users can control how to enhance the images. We show an example in Figure 5, where the input image has three different degradations, and we aim to focus on a particular one. Although these results do not offer great reconstruction, we believe it is a promising direction that illustrates the effectiveness of instruction guidance for image restoration and enhancement. We provide more results in Figures 6 and 7, where we show the potential of our method to restore and enhance images in a controllable manner.


Figure 6. **Instruction-based Image Restoration**. *InstructIR* understands a wide a range of instructions for a given task (first row). Given an *adversarial instruction* (second row), the model performs an identity –we did not enforce this during training–. Images from BSD68 [52].


Figure 7. **Multiple Real Instructions**. We can prompt multiple instructions (in sequence) to restore and enhance the images. This provides additional *control*. We show two examples of multiple instructions applied to the “Input” image -from left to right-.

Table 7. Quantitative comparisons with *state-of-the-art* methods on the **LOL dataset [83] (low-light enhancement)**. Table based on [81].

| **Method** | LPNet [43] | URetinex -Net[84] | DeepLPF [55] | SCI [51] | LIME [27] | MF [23] | NPE [78] | SRIE [24] | SDD [28] | CDEF [40] | ***InstructIR*** *Ours* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **PSNR** ↑ | 21.46 | 21.32 | 15.28 | 15.80 | 16.76 | 16.96 | 16.96 | 11.86 | 13.34 | 16.33 | <u>22.83</u> |
| **SSIM** ↑ | 0.802 | 0.835 | 0.473 | 0.527 | 0.444 | 0.505 | 0.481 | 0.493 | 0.635 | 0.583 | <u>0.836</u> |
| **Method** | DRBN [89] | KinD [107] | RUAS [46] | FIDE [86] | EG [33] | MS-RDN [90] | Retinex -Net[83] | MIRNet [93] | IPT [8] | Uformer [82] | **IAGC** [81] |
| **PSNR** ↑ | 20.13 | 20.87 | 18.23 | 18.27 | 17.48 | 17.20 | 16.77 | 24.14 | 16.27 | 16.36 | **24.53** |
| **SSIM** ↑ | 0.830 | 0.800 | 0.720 | 0.665 | 0.650 | 0.640 | 0.560 | 0.830 | 0.504 | 0.507 | **0.842** |

Table 8. Comparison with general restoration and all-in-one methods (*) at **image denoising**. We report PSNR on benchmark datasets considering different $\sigma$ noise levels. Table based on [100].

| | **CBSD68 [52]** | | | **Urban100 [32]** | | | **Kodak24 [22]** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **15** | **25** | **50** | **15** | **25** | **50** | **15** | **25** | **50** |
| IRCNN [103] | 33.86 | 31.16 | 27.86 | 33.78 | 31.20 | 27.70 | 34.69 | 32.18 | 28.93 |
| FFDNet [104] | 33.87 | 31.21 | 27.96 | 33.83 | 31.40 | 28.05 | 34.63 | 32.13 | 28.98 |
| DnCNN [101] | 33.90 | 31.24 | 27.95 | 32.98 | 30.81 | 27.59 | 34.60 | 32.14 | 28.95 |
| NAFNet [10] | 33.67 | 31.02 | 27.73 | 33.14 | 30.64 | 27.20 | 34.27 | 31.80 | 28.62 |
| HINet [9] | 33.72 | 31.00 | 27.63 | 33.49 | 30.94 | 27.32 | 34.38 | 31.84 | 28.52 |
| DGUNet [56] | 33.85 | 31.10 | 27.92 | 33.67 | 31.27 | 27.94 | 34.56 | 32.10 | 28.91 |
| MIRNetV2 [93] | 33.66 | 30.97 | 27.66 | 33.30 | 30.75 | 27.22 | 34.29 | 31.81 | 28.55 |
| SwinIR [44] | 33.31 | 30.59 | 27.13 | 32.79 | 30.18 | 26.52 | 33.89 | 31.32 | 27.93 |
| Restormer [95] | 34.03 | 31.49 | 28.11 | 33.72 | 31.26 | 28.03 | 34.78 | 32.37 | 29.08 |
| * DL [21] | 23.16 | 23.09 | 22.09 | 21.10 | 21.28 | 20.42 | 22.63 | 22.66 | 21.95 |
| * T.weather [75] | 31.16 | 29.00 | 26.08 | 29.64 | 27.97 | 26.08 | 31.67 | 29.64 | 26.74 |
| * TAPE [45] | 32.86 | 30.18 | 26.63 | 32.19 | 29.65 | 25.87 | 33.24 | 30.70 | 27.19 |
| * AirNet [42] | 33.49 | 30.91 | 27.66 | 33.16 | 30.83 | 27.45 | 34.14 | 31.74 | 28.59 |
| * IDR [100] | 34.11 | 31.60 | 28.14 | 33.82 | 31.29 | 28.07 | 34.78 | 32.42 | 29.13 |
| * *InstructIR*-**5D** | 34.00 | 31.40 | 28.15 | 33.77 | 31.40 | 28.13 | 34.70 | 32.26 | 29.16 |
| * *InstructIR*-**3D** | **34.15** | **31.52** | **28.30** | **34.12** | **31.80** | **28.63** | **34.92** | **32.50** | **29.40** |

Table 9. **Deblurring and Dehazing comparisons.** We compare with task-specific classical methods on benchmark datasets.

| **Deblurring GoPro [57]** | | **Dehazing SOTS [41]** | |
|---|---|---|---|
| **Method** | **PSNR/SSIM** | **Method** | **PSNR/SSIM** |
| Xu *et al.* [87] | 21.00/0.741 | DehazeNet [6] | 22.46/0.851 |
| DeblurGAN [38] | 28.70/0.858 | GFN [65] | 21.55/0.844 |
| Nah *et al.* [57] | 29.08/0.914 | GCANet [7] | 19.98/0.704 |
| RNN [99] | 29.19/0.931 | MSBDN [17] | 23.36/0.875 |
| DeblurGAN-v2 [39] | <u>29.55/0.934</u> | DuRN [47] | 24.47/0.839 |
| *InstructIR*-**5D** | 29.40/0.886 | *InstructIR*-**5D** | <u>27.10/0.956</u> |
| *InstructIR*-**6D** | **29.73/0.892** | *InstructIR*-**3D** | **30.22/0.959** |

This implies an advancement *w.r.t* classical (deterministic) image restoration methods. Classical deep restoration methods lead to a unique result, thus, they do not allow to control how the image is processed. We also compare *InstructIR* with InstructPix2Pix [4] in Figure 8.

**Qualitative Results.** We provide diverse qualitative results for several tasks. In Figure 9, we show results on the LOL [83] dataset. In Figure 10, we compare methods on the motion deblurring task using the GoPro [57] dataset. In Figure 11, we compare with different methods for the dehazing task on SOTS (outdoor) [41]. In Figure 12, we compare with image restoration methods for deraining on Rain100L [21]. Finally, we show denoising results in Figure 13. In this qualitative analysis, we use our single *InstructIR*-**5D** model to restore all the images.

**Discussion on Instruction-based Restoration** In Figure 8 we compare with InstructPix2Pix [4]. Our method is notably superior in terms of efficiency, fidelity and quality. We can conclude that diffusion-based methods [4, 53, 67] for image manipulation require complex "tuning" of several (hyper-)parameters, and strong regularization to enforce fidelity and reduce hallucinations. InstructPix2Pix [4] cannot solve inverse problems directly –although it has a good prior for solving Inpainting–, which indicates that such model require restoration-specific training (or fine-tuning).

**Limitations** Our method achieves *state-of-the-art* results in five tasks, proving the potential of using instructions to guide deep blind restoration models. However, we acknowledge certain limitations. First, in comparison to diffusion-based restoration methods, our current approach would not produce better results attending to perceptual quality. Second, our model struggles to process images with more than one degradation (*i.e. real-world* images), yet this is a common limitation among the related restoration methods. Third, as previous *all-in-one* methods, our model only works with *in-distribution degradations*, thus it will not work on unseen artifacts. Nevertheless, these limitations can be surpassed with more realistic training data.

## 7. Conclusion

We present the first approach that uses human-written instructions to guide the image restoration models. Given natural language prompts, our model can recover high-quality images from their degraded counterparts, considering multiple degradation types. InstructIR achieves state-of-the-art results on several restoration tasks, demonstrating the power of instruction guidance. These results represent a novel benchmark for text-guided image restoration.n

**Acknowledgments** This work was partly supported by the The Humboldt Foundation (AvH). Marcos Conde is also supported by Sony Interactive Entertainment, FTG.

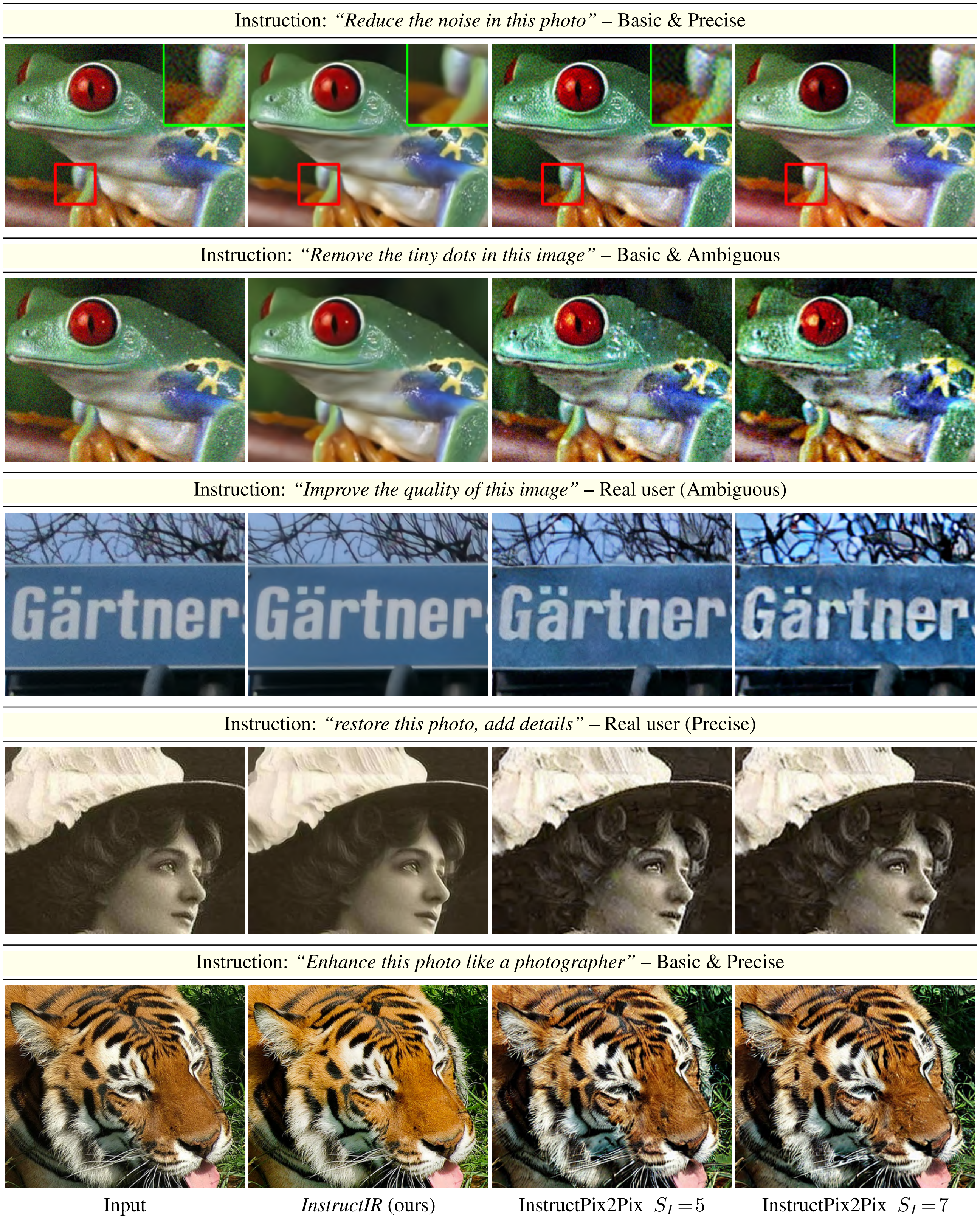



Figure 8. **Comparison with InstructPix2Pix [4] for instruction-based restoration using the prompt**. Images from the *RealSRSet* [44, 80]. We use our **7D** variant. We run InstructPix2Pix [4] using two configurations where we vary the weight of the image component hoping to improve fidelity: $S_I = 5$ and $S_I = 7$ (also known as Image CFG), this parameters helps to enforce fidelity and reduce hallucinations.

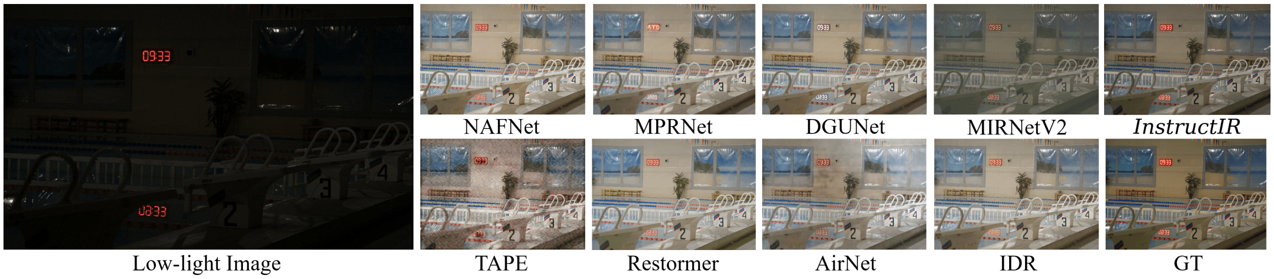



Figure 9. **Low-light Image Enhancement Results.** We compare with other methods on LOL [83] (`748.png`).

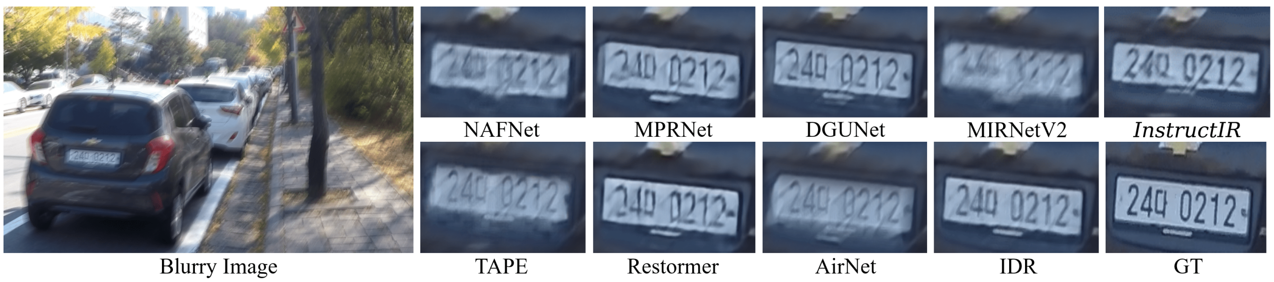



Figure 10. **Image Deblurring Results.** Comparison with other methods on the GoPro [57] dataset (`GOPR0854-11-00-000001.png`).

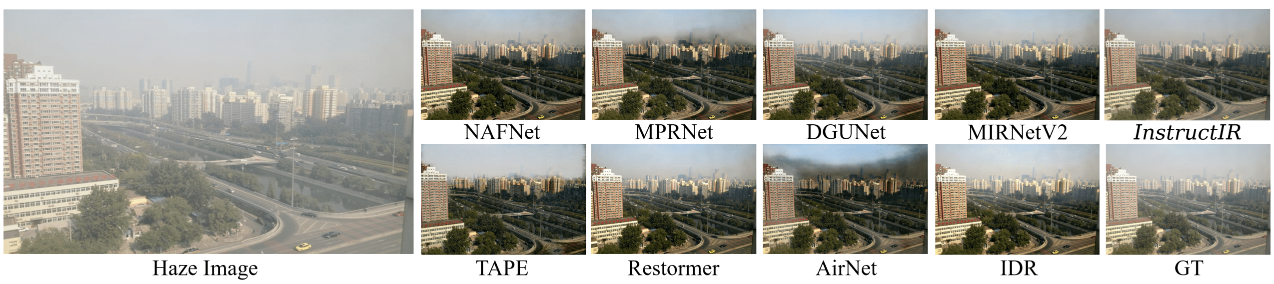



Figure 11. **Image Dehazing Results.** Comparison with other methods on SOTS [41] *outdoor* (`0150.jpg`).

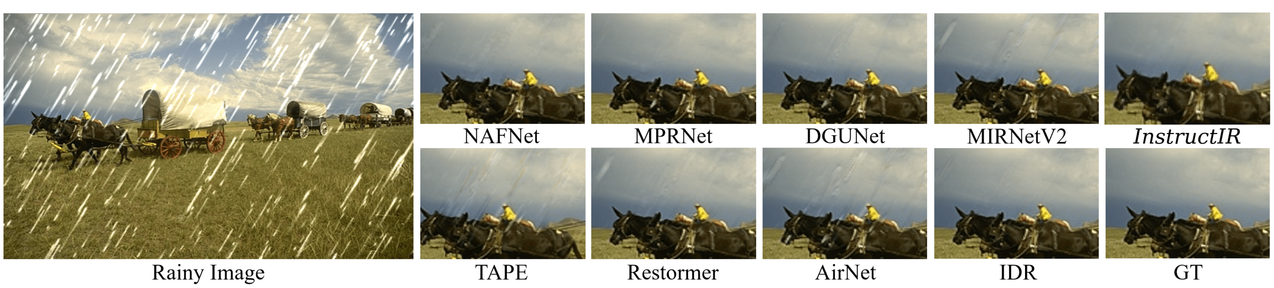



Figure 12. **Image Deraining Results** on Rain100L [21] (`035.png`).

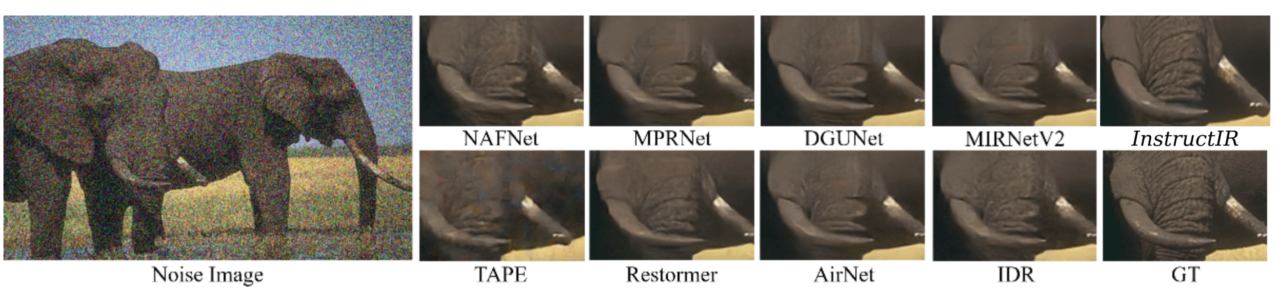



Figure 13. **Image Denoising Results** on BSD68 [52] (`0060.png`).

# High-Quality Image Restoration Following Human Instructions

## Supplementary Material

We define our loss functions in the paper *Sec. 4.1*. Our training loss function is $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{ce}$, which includes the loss function of the image model ($\mathcal{L}_1$), and the loss function for intent (task/degradation) classification ($\mathcal{L}_{ce}$) given the prompt embedding. We provide the loss evolution plots in Figures 14 and 15. In particular, in Figure 15 we can observe how the intent classification loss (*i.e.* predicting the task (or degradation) given the prompt), tends to 0 very fast, indicating that our language model component can infer easily the task given the instruction.

Additionally, we study three different text (sentence) encoders: (i) BGE-MICRO-V2 [2], (ii) ALL-MINILM-L6-V2 [3], (iii) CLIP text encoder (OpenAI CLIP ViT B-16). Note that these are always frozen. We use pre-trained weights from HuggingFace.

In Table 10 we show the ablation study. There is no significant difference between the text encoders. This is related to the previous results (Fig. 15), any text encoder with enough complexity can infer the task from the prompt. Therefore, we use BGE-MICRO-V2, as it is just 17M parameters in comparison to the others (40-60M parameters). *Note that for this ablation study, we keep fixed the image model (16M), and we only change the language model.*

**Text Discussion** We shall ask, *do the text encoders perform great because the language and instructions are too simple?*

We believe our instructions cover a wide range of expressions (technical, common language, ambiguous, etc). The language model works properly on real-world instructions. Therefore, we believe the language for this specific task is self-constrained, and easier to understand and to model in comparison to other "open" tasks such as image generation.

**Model Design** Based on our experiments, given a trained text-guided image model (*e.g.* based on NAFNet [10]), we can switch language models without performance loss.

*Comparison of NAFNet with and without using text (i.e. image only)*: The reader can find the comparison in the main paper Table 2, please read the highlighted caption.

*How the 6D variant does Super-Resolution?*. We degraded the input images by downsampling and re-upsampling using Bicubic interpolation. Given a LR image, we updample it using Bicubic, then InstructIR can recover some details.

Table 10. **Ablation study on the text encoders.** We report PSNR/SSIM metrics for each task using our **5D** base model. We use the same fixed image model (based on NAFNet [10]).

| Encoder | Deraining | Denoising | Deblurring | LOL |
|---|---|---|---|---|
| BGE-MICRO | 36.84/0.973 | 31.40/0.887 | 29.40/0.886 | 23.00/0.836 |
| ALL-MINILM | 36.82/0.972 | 31.39/0.887 | 29.40/0.886 | 22.98/0.836 |
| CLIP | 36.83/0.973 | 31.39/0.887 | 29.40/0.886 | 22.95/0.834 |

[2] https://huggingface.co/TaylorAI/bge-micro-v2

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

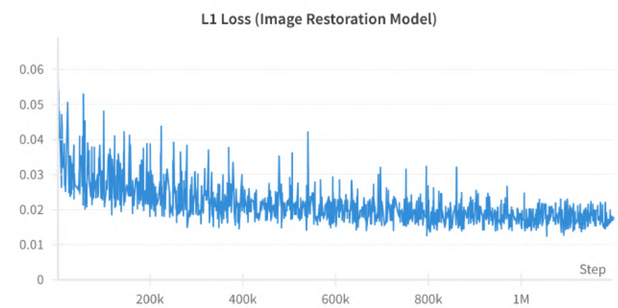



Figure 14. Image Restoration Loss ($\mathcal{L}_1$) computed between the restored image $\hat{x}$ (model's output) and the reference image $x$.

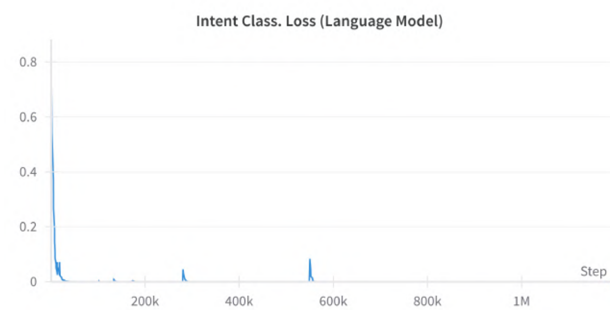
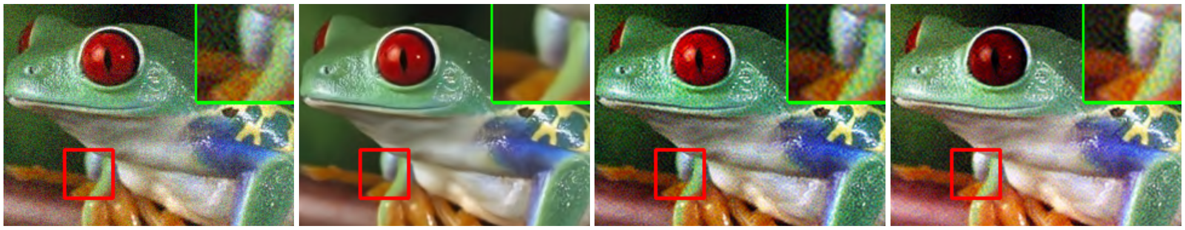



Figure 15. Intent Classification Loss from the instructions. Product of our simple MLP classification head using $\mathbf{e}$. When $\mathcal{L}_{ce} \to 0$ the model uses the learned (optimized) prompt embeddings, and it is optimized mainly based on the image regression loss ($\mathcal{L}_1$).

**Real-World Generalization.** We evaluate *InstructIR* as previous works [42, 61, 100]. Also, we find the same limitations as such methods when we process real-world images. Evaluating the model on (multiple) real-world degradations is a future task.

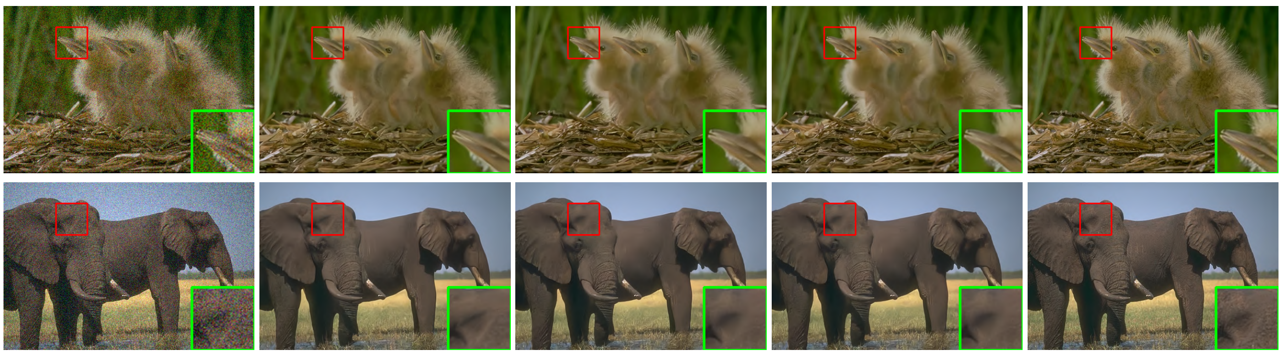
**Contemporary Works and Reproducibility.** Note that PromptIR, ProRes [49] and Amirnet [98] are contemporary works (presented or published by Dec 2023). We compare mainly with AirNet [42] since the model and results are open-source, and it is a reference all-in-one method. To the best of our knowledge, IDR [100] and ADMS [60] do not provide open-source code, models or results, thus we cannot compare with them qualitatively.

**Additional Visual Results** We present diverse qualitative samples in Figures 17, 18, and 19. Our method produces high-quality results given images with any of the studied degradations. In most cases the results are better than the reference all-in-one model AirNet [42]. Download all the test results at https://github.com/mv-lab/InstructIR.



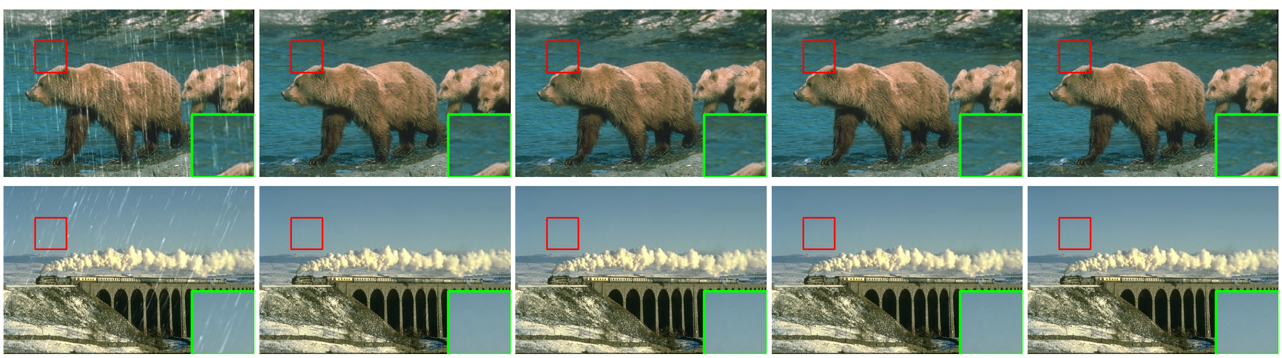
Input (RealSRSet) *InstructIR* InstructPix2Pix #1 InstructPix2Pix #2

Figure 16. **Comparison with InstructPix2Pix [4]** for instruction-based restoration using the prompt *"Remove the noise in this photo"*.



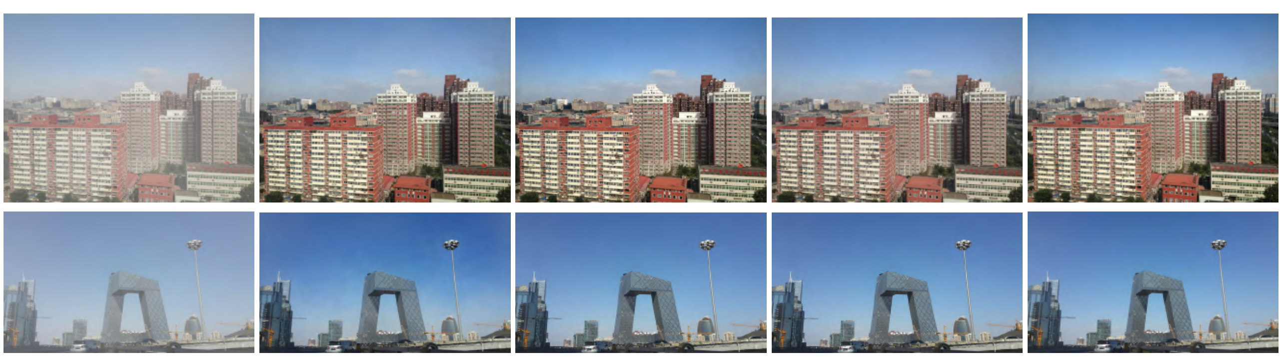
Input AirNet [42] PromptIR [61] *InstructIR* Reference

Figure 17. **Denoising results** for all-in-one methods. Images from BSD68 [52] with noise level $\sigma = 25$.



Input AirNet [42] PromptIR [61] *InstructIR* Reference

Figure 18. **Image deraining comparisons** for all-in-one methods on images from the Rain100L dataset [21].



Input AirNet [42] PromptIR [61] *InstructIR* Reference

Figure 19. **Dehazing comparisons** for all-in-one methods on images from the SOTS outdoor dataset [41].

## References


[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017. 6

[2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2011. 5

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

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402. IEEE, 2023. 2, 3, 9, 10, 13

[5] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 5, 7

[6] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 9

[7] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1375–1383. IEEE, 2019. 9

[8] 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*, 2021. 9

[9] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 6, 9

[10] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 1, 2, 4, 5, 6, 9, 12

[11] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6306–6314, 2018. 7

[12] Victor Cornillere, Abdelaziz Djelouah, Wang Yifan, Olga Sorkine-Hornung, and Christopher Schroers. Blind image super-resolution with spatially variant degradations. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 5

[14] Chuntao Ding, Zhichao Lu, Shangguang Wang, Ran Cheng, and Vishnu Naresh Boddeti. Mitigating task interference in multi-task learning via explicit task routing with nonlearnable primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7756–7765, 2023. 4

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2015. 6

[16] 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*, 2020. 1, 2

[17] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2157–2167, 2020. 9

[18] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *TIP*, 2011. 2

[19] Yu Dong, Yihao Liu, He Zhang, Shifeng Chen, and Yu Qiao. Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10729–10736, 2020. 7

[20] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997. 1

[21] Qingnan Fan, Dongdong Chen, Lu Yuan, Gang Hua, Nenghai Yu, and Baoquan Chen. A general decoupled learning framework for parameterized image operators. *IEEE transactions on pattern analysis and machine intelligence*, 43 (1):33–47, 2019. 6, 7, 9, 11, 13

[22] Rich Franzen. Kodak lossless true color image suite. http://r0k.us/graphics/kodak/, 1999. Online accessed 24 Oct 2021. 9

[23] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. 129:82–96, 2016. 9

[24] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, 2016. 9

[25] 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. 7

[26] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral

learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 7

[27] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016. 9

[28] Shijie Hao, Xu Han, Yanrong Guo, Xin Xu, and Meng Wang. Low-light image enhancement with semi-decoupled decomposition. *IEEE TMM*, 22(12):3025–3038, 2020. 9

[29] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 2010. 2

[30] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[31] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, 2019. 2

[32] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 5, 9

[33] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 30:2340–249, 2021. 9

[34] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[35] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *TPAMI*, 2010. 2

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[37] Johannes Kopf, Boris Neubert, Billy Chen, Michael Cohen, Daniel Cohen-Or, Oliver Deussen, Matt Uyttendaele, and Dani Lischinski. Deep photo: Model-based photograph enhancement and viewing. *ACM TOG*, 2008. 2

[38] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. DeblurGAN: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 9

[39] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 9

[40] Xiaozhou Lei, Zixiang Fei, Wenju Zhou, Huiyu Zhou, and Minrui Fei. Low-light image enhancement using the cell vibration model. *IEEE TMM*, pages 1–1, 2022. 9

[41] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 5, 6, 7, 9, 11, 13

[42] 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, 5, 6, 7, 9, 12, 13

[43] Jiaqian Li, Juncheng Li, Faming Fang, Fang Li, and Guixu Zhang. Luminance-aware pyramid network for low-light image enhancement. *IEEE TMM*, 23:3153–3165, 2020. 9

[44] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV Workshops*, 2021. 1, 2, 6, 9, 10

[45] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 447–464. Springer, 2022. 6, 9

[46] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 9

[47] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 9

[48] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Learning the degradation distribution for blind image super-resolution. In *CVPR*, pages 6063–6072, 2022. 2

[49] Jiaqi Ma, Tianheng Cheng, Guoli Wang, Qian Zhang, Xinggang Wang, and Lefei Zhang. Prores: Exploring degradation-aware visual prompt for universal image restoration. *arXiv preprint arXiv:2306.13653*, 2023. 2, 12

[50] 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*, 2016. 5

[51] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 9

[52] 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*, 2001. 5, 6, 7, 8, 9, 11, 13

[53] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 9

[54] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *ICCV*, 2013. 2

[55] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *CVPR*, 2020. 9

[56] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17399–17410, 2022. 6, 9

[57] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 5, 6, 9, 11

[58] Seungjun Nah, Sanghyun Son, Jaerin Lee, and Kyoung Mu Lee. Clean images are hard to reblur: Exploiting the ill-posed inverse task for dynamic scene deblurring. In *ICLR*, 2022. 1, 2

[59] Nhat Nguyen, Peyman Milanfar, and Gene Golub. Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. *IEEE Transactions on image processing*, 10(9):1299–1308, 2001. 1

[60] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5815–5824. IEEE, 2023. 2, 5, 12

[61] 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. 2, 5, 7, 12, 13

[62] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3

[63] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019. 3

[64] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *CVPR*, 2021. 1, 2

[65] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3253–3261, 2018. 9

[66] Wenqi Ren, Jinshan Pan, Hua Zhang, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks with holistic edges. *IJCV*, 2020. 1

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 3, 9

[68] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[69] Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*, 2017. 4

[70] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[71] Gjorgji Strezoski, Nanne van Noord, and Marcel Worring. Many task learning with task routing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1375–1384, 2019. 4, 5

[72] Chunwei Tian, Yong Xu, and Wangmeng Zuo. Image denoising using deep cnn with batch renormalization. *Neural Networks*, 2020. 7

[73] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, 2013. 2

[74] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MAXIM: Multi-axis MLP for image processing. In *CVPR*, pages 5769–5780, 2022. 1, 2, 5

[75] 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, 6, 9

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[77] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6849–6857, 2019. 7

[78] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE TIP*, 22(9):3538–3548, 2013. 9

[79] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[80] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 10

[81] Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, and Shuaicheng Liu. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13128–13137, 2023. 9

[82] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv:2106.03106*, 2021. 1, 2, 9

[83] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018. 5, 6, 9, 11

[84] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *CVPR*, 2022. 9

[85] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597, 2023. 5

[86] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, 2020. 9

[87] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *CVPR*, 2013. 9

[88] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 5, 6

[89] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. Band representation-based semi-supervised low-light image enhancement: bridging the gap between signal fidelity and perceptual quality. *IEEE TIP*, 30:3461–3473, 2021. 9

[90] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE TIP*, 30:2072–2086, 2021. 9

[91] Mingde Yao, Ruikang Xu, Yuanshen Guan, Jie Huang, and Zhiwei Xiong. Neural degradation representation learning for all-in-one image restoration. *arXiv preprint arXiv:2310.12848*, 2023. 2

[92] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 2

[93] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020. 6, 9

[94] 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*, 2021. 7

[95] 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*, 2022. 1, 2, 6, 9

[96] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (4):2058–2073, 2020. 7

[97] Cheng Zhang, Yu Zhu, Qingsen Yan, Jinqiu Sun, and Yanning Zhang. All-in-one multi-degradation image restoration network via hierarchical degradation representation. *arXiv preprint arXiv:2308.03021*, 2023. 2

[98] Cheng Zhang, Yu Zhu, Qingsen Yan, Jinqiu Sun, and Yanning Zhang. All-in-one multi-degradation image restoration network via hierarchical degradation representation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2285–2293, 2023. 2, 12

[99] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, 2018. 9

[100] Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5825–5835, 2023. 2, 5, 6, 9, 12

[101] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 2017. 9

[102] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1

[103] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, 2017. 1, 9

[104] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *TIP*, 2018. 9

[105] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, pages 2737–2746, 2020. 1

[106] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[107] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, 2019. 9