InstructIR: High-Quality Image Restoration Following Human Instructions

Marcos V. Conde[1,2], Gregor Geigle[1], and Radu Timofte[1]

[1] Computer Vision Lab, CAIDAS & IFI, University of Würzburg
[2] Sony PlayStation, FTG
https://github.com/mv-lab/InstructIR

Fig. 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 in some *real-world* scenarios (last three images, from left to right).

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 [13, 16, 38, 52, 90, 91].

Some methods focus on specific degradations, for instance reducing noise (denoising) [57, 90, 91], removing blur (deblurring) [51, 92], or clearing haze (dehazing) [13, 58]. 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 [7, 66, 73, 83], yet training the neural network for each specific task independently. Since using a separate model for each possible degradation is resource-intensive, recent approaches propose *All-in-One* restoration models [36, 53, 54, 89]. These approaches use a single deep blind restoration model considering multiple degradation types and levels. Contemporary works such as PromptIR [54] or ProRes [42] 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 [36,53,89]. 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 multi-task 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 complex 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 [13, 38, 51, 57, 66, 83] have shown consistently better results compared to traditional techniques for blind image restoration [14, 25, 31, 33, 47, 65]. The proposed neural networks are based on convolutional neural networks (CNNs) and Transformers [68] (or related attention mechanisms). We focus on general-purpose restoration models [7, 38, 73, 83]. For example, SwinIR [38], MAXIM [66] and Uformer [73]. 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 [83] uses non-local blocks [71] to capture complex features across the image.

NAFNet [7] 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 building block (NAFBlock) follows a simple metaformer [82] architecture with efficient inverted residual blocks [27]. 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 selecting 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 [36,42,53,54,67,81,87,88]. These approaches use a single deep blind restoration model to tackle different degradation types and levels.

We use as reference AirNet [36], IDR [89] and ADMS [53]. We also consider the contemporary work PromptIR [54]. The methods use different techniques to guide the blind model in the restoration process. For instance, an auxiliary model for degradation classification [36,53], or multi-dimensional guidance vectors (also known as "prompts") [41,42,54] that help the model to discriminate the different types of degradation in the image.

Text-guided Image Manipulation. In recent years, multiple methods have been proposed for text-to-image generation and text-based image editing works [4, 26, 30, 46, 62]. 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.

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. We also consider the language complexity, from ambiguous instructions (*e.g.* "fix my image") to precise instructions (*e.g.* "remove the noise").

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.

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

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

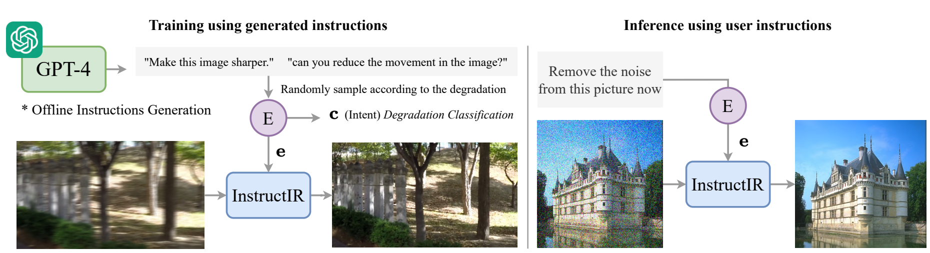


Fig. 2: We train our blind image restoration models using common image datasets, and prompts generated using GPT-4, note that this is (self-)supervised learning. At inference time, our model generalizes to human-written instructions and restores (or enhances) the images.

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 [59] and manipulation [3, 4] often use the text encoder of a CLIP model [55] 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). We opt, instead, to use a pure text-based sentence encoder [56], that is, a smaller 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. We compare text encoders in the supplementary material.

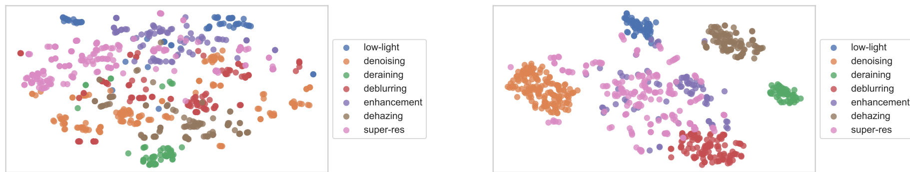
Fine-tuning the Text Encoder. We want to adapt the text encoder E for the restoration task to better encode the required information for the restoration model. Training the full text encoder is likely to lead to overfitting on our small training set and lead to loss of generalization. Instead, we freeze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

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

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 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 denoising. In our experiments, $d_t = 384$, $d_v = 256$ and $\mathbf{W}$ is a linear layer. The representation $\mathbf{e}$ from the text encoder is shared across the blocks, and each block has a trainable projection $\mathbf{W}$.

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 degradation classes. The classification head $\mathcal{C}$ is a simple two-layers MLP. Thus, we only need to train a projection layer $\mathbf{W}$ 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.



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

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

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

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 [7] as the image model, an efficient image restoration model that follows a U-Net architecture [60]. To successfully learn multiple tasks using a single model, we use task routing techniques. Our framework for training and evaluation 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 [11, 61, 63], we propose an *"Instruction Condition Block" (ICB)* to enable task-specific transformations within the model. Conventional task routing [63] 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.

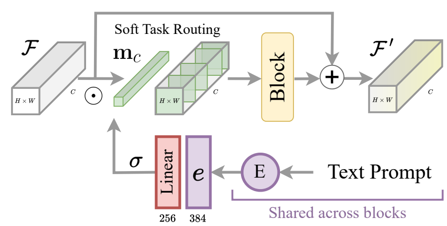


Fig. 4: *Instruction Condition Block (ICB)* using an approximation of task routing [63] for many-tasks learning (See Eq. 2). This mechanism allows the neural network to select and prioritize specific features depending on the instruction, similarly to a Mixture of Experts (MoE).

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

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

where the mask $\mathbf{m}_c = \sigma(\mathbf{W_c} \cdot \mathbf{e})$ is produced using a linear layer $\mathbf{W_c}$ – 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 [24, 63], 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 convolutional NAFBlock [7] (Block). We illustrate our task-routing ICB block in Figure 4. We use "regular" NAFBlocks [7], followed by ICBs to condition the features, at both encoder and decoder blocks. The formulation is $F^{l+1} = \text{ICB}(\text{Block}(F^l))$ where l is the layer. Although we do not condition explicitly the filters of the neural network, as in [63], 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 [11,63].

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 [36,53].

4 Experimental Results

We evaluate our model on 9 well-known benchmarks for different image restoration tasks: image denoising, deblurring, deraining, dehazing, real low-light enhancement, and photo-realistic image enhancement. We present extensive quantitative results in Table 2 and Table 3. We provide extensive comparisons with other all-in-one methods as well as task-specific methods. 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 but 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 [3], a distilled version of BGE-small-en [76]. The BGE encoders are BERT-like encoders [10] 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 [7] as the image model backbone. 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

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

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 [63] 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 using a batch size of 32 and AdamW [32] 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 256×256 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 (>10K 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 (*e.g.* Google Colab T4 16Gb GPU).

4.2 Datasets and Benchmarks

Following previous works [36,54,89], we prepare the datasets for different restoration tasks, including real and synthetic datasets.
Image denoising. We use a combination of BSD400 [2] and WED [43] 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 [45] and Urban100 [28] datasets.
Image deraining. We use the Rain100L [78] dataset, which consists of 200 clean-rainy image pairs for training, and 100 pairs for testing.
Image dehazing. We utilize the Reside (outdoor) SOTS [35] 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* test set of 500 images.
Image deblurring. We use the GoPro dataset for motion deblurring [50] which consists of 2103 images for training, and 1111 for testing.
Real-world Low-light Image Enhancement. We use the LOL [74] dataset (v1), which contains real-case low/normal-light image pairs. We adopt its official split of 485 training images and 15 testing images.
Real-world Image Enhancement. Extending previous works, we also study photo-realistic image enhancement using the MIT5K DSLR dataset [5]. We use 1000 images for training, and the standard split of 500 images for testing (as in [66]).

Finally, as previous works [36,54,89], we combine all the aforementioned training datasets, and we train our unified model for all-in-one restoration. Note that we do not include more *real-world datasets* because previous works do not provide results (or models) for those. Moreover, previous works were limited to synthetic data, in contrast, *InstructIR* also tackles real-world image enhancement.

Table 2: Quantitative results on **five restoration tasks (5D)** with *state-of-the-art* general image restoration and all-in-one methods. We highlight the reference model *without* text (image only), the best overall 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 [89].

Methods	Deraining Rain100L [78]		Dehazing SOTS [35]		Denoising BSD68 [45]		Deblurring GoPro [50]		Low-light Enh. LOL [74]		Average		Params
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	(M)
HINet [8]	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 [49]	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 [84]	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 [38]	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 [83]	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 [7]	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 [17]	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 [67]	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 [39]	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 [36]	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
InstructIR w/o text	35.58	0.967	25.20	0.938	31.09	0.883	26.65	0.810	20.70	0.820	27.84	0.884	17.11
IDR [89]	35.63	0.965	25.24	0.943	31.60	0.887	27.87	0.846	21.34	0.826	28.34	0.893	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

Table 3: Comparisons of all-in-one restoration models for **3 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 [54].

Methods	Dehazing SOTS [35]	Deraining Rain100L [17]	Denoising ablation study (BSD68 [45]) $\sigma = 15$	$\sigma = 25$	$\sigma = 50$	Average
BRDNet [64]	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 [20]	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 [15]	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 [85]	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 [17]	26.92/0.931	32.62/0.931	33.05/0.914	30.41/0.861	26.90/0.740	29.98/0.875
AirNet [36]	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 [54]	**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
InstructIR w/o text	26.84/0.948	34.02/0.960	33.70/0.929	30.94/0.882	27.78/0.780	30.65/0.900

4.3 Multiple Degradation Results

We define two initial setups for multi-task restoration:

- **3D** for *three-degradation* models such as AirNet [36], 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 [89].

In Table 2, we show the performance of **5D** models. Following Zhang *et al.* [89], we compare *InstructIR* with several *state-of-the-art* methods for general image restoration [7,8,38,83,84], and all-in-one image restoration methods [17,36, 39,67,89]. 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 3 we can appreciate a similar behavior, 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 and the human-written instruction. Therefore, we also prepare a test set 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 4. *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 Figures 5 and 6.

Table 4: 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 test set 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 Discussion and 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 [1,12] 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.

How does 7D work? Finally, if we add real 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 on this task using the MIT5K [5] dataset is notable, while keeping the performance on the other tasks.

Table 5: Real Image Enhancement results on MIT5K [5].

Method	PSNR ↑	SSIM ↑	ΔE_{ab} ↓
UPE [69]	21.88	0.853	10.80
DPE [21]	23.75	0.908	9.34
HDRNet [9]	24.32	<u>0.912</u>	8.49
3DLUT [86]	**25.21**	**0.922**	**7.61**
InstructIR-**7D**	<u>24.65</u>	0.900	<u>8.20</u>

Fig. 5: Adversarial and OOD samples for Instruction-based Restoration. *InstructIR* understands a wide range of instructions for a given task (first row). Given an *adversarial or out-of-distribution instruction* (second row), the model does not modify the image notably (*i.e.* performs the identity) –we did not enforce this during training–.

Table 7: Quantitative comparisons with *state-of-the-art* methods on the **LOL [74]** dataset for Real-World Low-light Enhancement. Note that *InstructIR*-**7D** is a multi-task method, while the other methods are task-specific. Table based on [72].

Method	LPNet [37]	URetinex -Net [75]	DeepLPF [48]	SCI [44]	LIME [22]	MF [18]	NPE [70]	SRIE [19]	SDD [23]	CDEF [34]	***InstructIR*** *Ours*
PSNR ↑	21.46	21.32	15.28	15.80	16.76	16.96	16.96	11.86	13.34	16.33	22.81
SSIM ↑	0.802	0.835	0.473	0.527	0.444	0.505	0.481	0.493	0.635	0.583	0.836
Method	DRBN [79]	KinD [93]	RUAS [40]	FIDE [77]	EG [29]	MS-RDN [80]	Retinex -Net [74]	MIRNet [84]	IPT [6]	Uformer [73]	**IAGC** [72]
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**

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.

Table 6: Summary ablation study on the multi-task variants of *InstructIR* that tackle from 3 to 7 tasks.

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

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 real-world photography enhancement in Table 5, and for real-world low-light image enhancement in Table 7.

Fig. 6: Control via 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-.

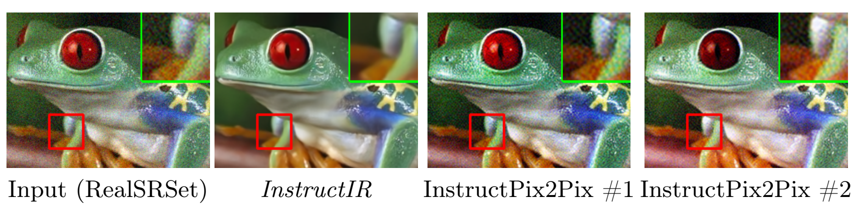


Fig. 7: Comparison with InstructPix2Pix [4] using the prompt *"Remove the noise in this photo"*. Real-case image from *RealSRSet* [38].

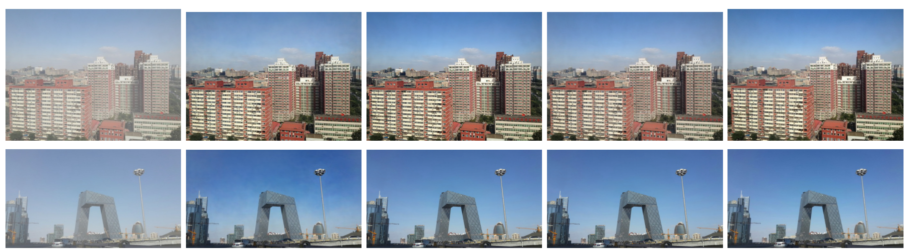
5.1 On the Effectiveness of Instructions

Thanks to our integration of human instructions, users can control how to enhance the images. We provide examples in Figures 5 and 6, where we show the potential of our method to restore and enhance images in a controllable manner.

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] (a diffusion-based generative model) in Figure 7.

Qualitative Results. We provide diverse qualitative results for several tasks, and we compare with all-in-one and task-specific methods. In Figure 10, we show

Fig. 8: Real-world samples of image restoration and enhancement using *InstructIR*.



Input | AirNet [36] | PromptIR [54] | *InstructIR* | Reference

Fig. 9: Dehazing comparisons for all-in-one restoration methods on SOTS [35].

results on the LOL [74] dataset. In Figure 11, we compare methods on the motion deblurring task using the GoPro [50] dataset. In Figures 9 and 12, we compare with different methods for the dehazing task on SOTS (outdoor) [35]. In Figure 13, we compare with image restoration methods for deraining on Rain100L [17]. Finally, we show denoising results in Figure 14. In this qualitative analysis, we use our single *InstructIR*-**5D** model to restore all the images.

Limitations As with previous *all-in-one* methods, our model struggles to process images with more than one degradation (*i.e.* complex *real-world* images), or unknown out-of-distribution degradations, yet this is a common limitation among the related methods. However, we believe that these limitations can be surpassed with more realistic training data, and scaling the model's complexity.

6 Conclusion

We present a novel approach that uses natural human-written instructions to guide the image restoration model. Given a prompt, our multi-task model can recover high-quality images from their degraded counterparts, considering multiple degradations. We achieve state-of-the-art results on several restoration tasks, demonstrating the power and flexibility of instruction guidance. Our results represent a new benchmark for text-guided image restoration.

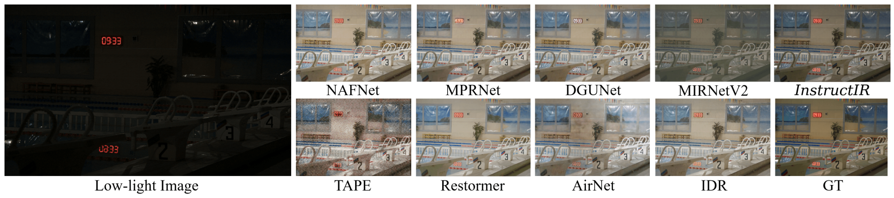


Fig. 10: Real Low-light Enhancement Results. LOL [74] testset (`748.png`). AirNet [36] and IDR [89] are well-known all-in-one restoration methods. NAFNet [7] is equivalent to *InstructIR* without text conditions (*i.e.* our image-only backbone).

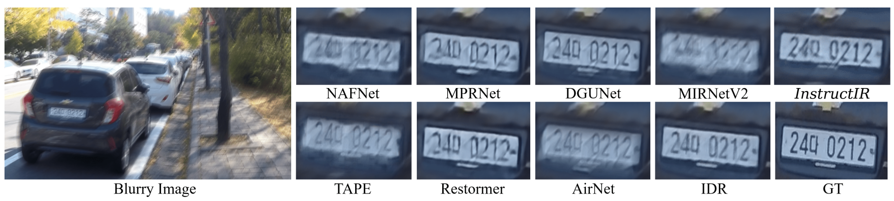


Fig. 11: Image Deblurring Results. GoPro [50] dataset.

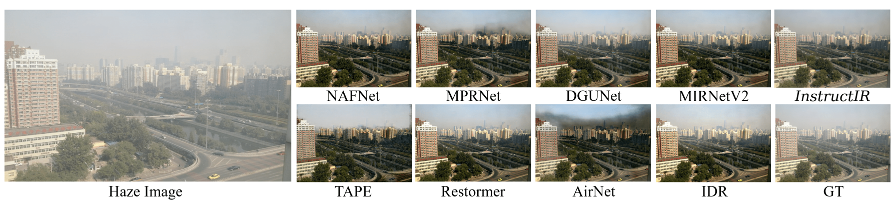


Fig. 12: Image Dehazing Results. SOTS [35] *outdoor* dataset (`0150.jpg`).

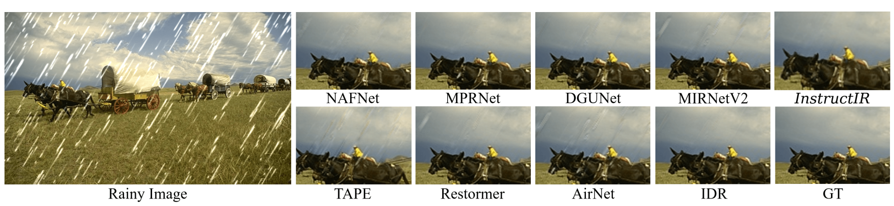


Fig. 13: Image Deraining Results on Rain100L [17] (`035.png`).

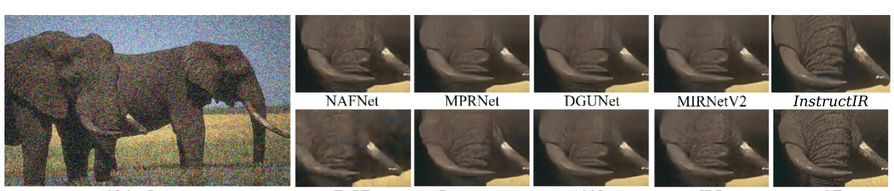


Fig. 14: Image Denoising Results on BSD68 [45] (`0060.png`).

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: CVPR Workshops (2017) 10
2. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. TPAMI (2011) 8
3. Bai, Y., Wang, C., Xie, S., Dong, C., Yuan, C., Wang, Z.: Textir: A simple framework for text-based editable image restoration. CoRR **abs/2302.14736** (2023). https://doi.org/10.48550/ARXIV.2302.14736, https://doi.org/10.48550/arXiv.2302.14736 5
4. Brooks, T., Holynski, A., Efros, A.A.: 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. pp. 18392–18402. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01764, https://doi.org/10.1109/CVPR52729.2023.01764 2, 3, 4, 5, 12
5. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: 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) 8, 10
6. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: CVPR (2021) 11
7. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV (2022) 2, 3, 6, 7, 9, 14
8. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 182–192 (2021) 9
9. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6306–6314 (2018) 10
10. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/N19-1423, https://doi.org/10.18653/v1/n19-1423 7
11. Ding, C., Lu, Z., Wang, S., Cheng, R., Boddeti, V.N.: Mitigating task interference in multi-task learning via explicit task routing with non-learnable primitives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7756–7765 (2023) 6, 7
12. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI (2015) 10
13. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: Multi-scale boosted dehazing network with dense feature fusion. In: CVPR (2020) 2
14. Dong, W., Zhang, L., Shi, G., Wu, X.: Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. TIP (2011) 2
15. Dong, Y., Liu, Y., Zhang, H., Chen, S., Qiao, Y.: Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10729–10736 (2020) 9

16. Elad, M., Feuer, A.: Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. IEEE transactions on image processing **6**(12), 1646–1658 (1997) 2
17. Fan, Q., Chen, D., Yuan, L., Hua, G., Yu, N., Chen, B.: A general decoupled learning framework for parameterized image operators. IEEE transactions on pattern analysis and machine intelligence **43**(1), 33–47 (2019) 9, 13, 14
18. Fu, X., Zeng, D., Huang, Y., Liao, Y., Ding, X., Paisley, J.: A fusion-based enhancing method for weakly illuminated images **129**, 82–96 (2016) 11
19. Fu, X., Zeng, D., Huang, Y., Zhang, X.P., Ding, X.: A weighted variational model for simultaneous reflectance and illumination estimation. In: CVPR (2016) 11
20. Gao, H., Tao, X., Shen, X., Jia, J.: Dynamic scene deblurring with parameter selective sharing and nested skip connections. In: CVPR. pp. 3848–3856 (2019) 9
21. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017) 10
22. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE TIP **26**(2), 982–993 (2016) 11
23. Hao, S., Han, X., Guo, Y., Xu, X., Wang, M.: Low-light image enhancement with semi-decoupled decomposition. IEEE TMM **22**(12), 3025–3038 (2020) 11
24. He, J., Dong, C., Qiao, Y.: Modulating image restoration with continual levels via adaptive feature modification layers (2019), https://arxiv.org/abs/1904.08118 6
25. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. TPAMI (2010) 2
26. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 3
27. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: ICCV (2019) 3
28. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5197–5206 (2015) 8
29. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE TIP **30**, 2340–249 (2021) 11
30. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023) 3
31. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. TPAMI (2010) 2
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014) 8
33. Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., Lischinski, D.: Deep photo: Model-based photograph enhancement and viewing. ACM TOG (2008) 2
34. Lei, X., Fei, Z., Zhou, W., Zhou, H., Fei, M.: Low-light image enhancement using the cell vibration model. IEEE TMM pp. 1–1 (2022) 11
35. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2018) 8, 9, 13, 14

36. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: CVPR. pp. 17452–17462 (June 2022) 2, 3, 7, 8, 9, 13, 14
37. Li, J., Li, J., Fang, F., Li, F., Zhang, G.: Luminance-aware pyramid network for low-light image enhancement. IEEE TMM **23**, 3153–3165 (2020) 11
38. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: ICCV Workshops (2021) 2, 9, 12
39. Liu, L., Xie, L., Zhang, X., Yuan, S., Chen, X., Zhou, W., Li, H., Tian, Q.: 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. pp. 447–464. Springer (2022) 9
40. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: CVPR (2021) 11
41. Liu, Y., Liu, A., Gu, J., Zhang, Z., Wu, W., Qiao, Y., Dong, C.: Discovering distinctive "semantics" in super-resolution networks (2022), `https://arxiv.org/abs/2108.00406` 3
42. Ma, J., Cheng, T., Wang, G., Zhang, Q., Wang, X., Zhang, L.: Prores: Exploring degradation-aware visual prompt for universal image restoration. arXiv preprint arXiv:2306.13653 (2023) 2, 3
43. Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., Zhang, L.: Waterloo exploration database: New challenges for image quality assessment models. TIP (2016) 8
44. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: CVPR (2022) 11
45. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001) 8, 9, 14
46. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021) 3
47. Michaeli, T., Irani, M.: Nonparametric blind super-resolution. In: ICCV (2013) 2
48. Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G.: Deeplpf: Deep local parametric filters for image enhancement. In: CVPR (2020) 11
49. Mou, C., Wang, Q., Zhang, J.: Deep generalized unfolding networks for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17399–17410 (2022) 9
50. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017) 8, 9, 13, 14
51. Nah, S., Son, S., Lee, J., Lee, K.M.: Clean images are hard to reblur: Exploiting the ill-posed inverse task for dynamic scene deblurring. In: ICLR (2022) 2
52. Nguyen, N., Milanfar, P., Golub, G.: Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. IEEE Transactions on image processing **10**(9), 1299–1308 (2001) 2
53. Park, D., Lee, B.H., Chun, S.Y.: 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). pp. 5815–5824. IEEE (2023) 2, 3, 7
54. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.S.: Promptir: Prompting for all-in-one blind image restoration. arXiv preprint arXiv:2306.13090 (2023) 2, 3, 8, 9, 13

55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html 5
56. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) 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. pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/D19-1410, https://doi.org/10.18653/v1/D19-1410 5
57. Ren, C., He, X., Wang, C., Zhao, Z.: Adaptive consistency prior based deep network for image denoising. In: CVPR (2021) 2
58. Ren, W., Pan, J., Zhang, H., Cao, X., Yang, M.H.: Single image dehazing via multi-scale convolutional neural networks with holistic edges. IJCV (2020) 2
59. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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. pp. 10674–10685. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.org/10.1109/CVPR52688.2022.01042 5
60. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI (2015) 6
61. Rosenbaum, C., Klinger, T., Riemer, M.: Routing networks: Adaptive selection of non-linear functions for multi-task learning. arXiv preprint arXiv:1711.01239 (2017) 6
62. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023) 3
63. Strezoski, G., Noord, N.v., Worring, M.: Many task learning with task routing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1375–1384 (2019) 6, 7, 8
64. Tian, C., Xu, Y., Zuo, W.: Image denoising using deep cnn with batch renormalization. Neural Networks (2020) 9
65. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: ICCV (2013) 2
66. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: MAXIM: Multi-axis MLP for image processing. In: CVPR. pp. 5769–5780 (2022) 2, 8
67. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: CVPR. pp. 2353–2363 (2022) 3, 9
68. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 2
69. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6849–6857 (2019) 10

70. Wang, S., Zheng, J., Hu, H.M., Li, B.: Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE TIP **22**(9), 3538–3548 (2013) 11
71. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018) 3
72. Wang, Y., Liu, Z., Liu, J., Xu, S., Liu, S.: 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. pp. 13128–13137 (2023) 11
73. Wang, Z., Cun, X., Bao, J., Liu, J.: Uformer: A general u-shaped transformer for image restoration. arXiv:2106.03106 (2021) 2, 11
74. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018) 8, 9, 11, 13, 14
75. Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., Jiang, J.: Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In: CVPR (2022) 11
76. Xiao, S., Liu, Z., Zhang, P., Muennighof, N.: C-pack: Packaged resources to advance general chinese embedding. CoRR **abs/2309.07597** (2023). https://doi.org/10.48550/ARXIV.2309.07597, https://doi.org/10.48550/arXiv.2309.07597 7
77. Xu, K., Yang, X., Yin, B., Lau, R.W.: Learning to restore low-light images via decomposition-and-enhancement. In: CVPR (2020) 11
78. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: CVPR (2020) 8, 9
79. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: Band representation-based semi-supervised low-light image enhancement: bridging the gap between signal fidelity and perceptual quality. IEEE TIP **30**, 3461–3473 (2021) 11
80. Yang, W., Wang, W., Huang, H., Wang, S., Liu, J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. IEEE TIP **30**, 2072–2086 (2021) 11
81. Yao, M., Xu, R., Guan, Y., Huang, J., Xiong, Z.: Neural degradation representation learning for all-in-one image restoration. arXiv preprint arXiv:2310.12848 (2023) 3
82. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10819–10829 (2022) 3
83. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022) 2, 3, 9
84. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: ECCV (2020) 9, 11
85. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021) 9
86. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: 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) 10
87. Zhang, C., Zhu, Y., Yan, Q., Sun, J., Zhang, Y.: All-in-one multi-degradation image restoration network via hierarchical degradation representation. arXiv preprint arXiv:2308.03021 (2023) 3

88. Zhang, C., Zhu, Y., Yan, Q., Sun, J., Zhang, Y.: All-in-one multi-degradation image restoration network via hierarchical degradation representation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2285–2293 (2023) 3
89. Zhang, J., Huang, J., Yao, M., Yang, Z., Yu, H., Zhou, M., Zhao, F.: Ingredient-oriented multi-degradation learning for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5825–5835 (2023) 2, 3, 8, 9, 14
90. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017) 2
91. Zhang, K., Zuo, W., Gu, S., Zhang, L.: Learning deep CNN denoiser prior for image restoration. In: CVPR (2017) 2
92. Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., Li, H.: Deblurring by realistic blurring. In: CVPR. pp. 2737–2746 (2020) 2
93. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: ACM MM (2019) 11