InstructIR: High-Quality Image Restoration Following Human Instructions Supplementary Material

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

A Additional Training Details and Ablations

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 1 and 2. In particular, in Figure 2 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 [3], (ii) all-MiniLM-L6-v2 [4], (iii) CLIP text encoder (OpenAI CLIP ViT B-16). Note that these are always frozen. We use pre-trained weights from HuggingFace.

In Table 1 we show the ablation study. There is no significant difference between the text encoders. This is related to the previous results (Fig. 2), 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.

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

Model Design Based on our experiments, given a trained text-guided image model (*e.g.* based on NAFNet [4]), we can switch language models without performance loss.

Table 1: 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 [4]).

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

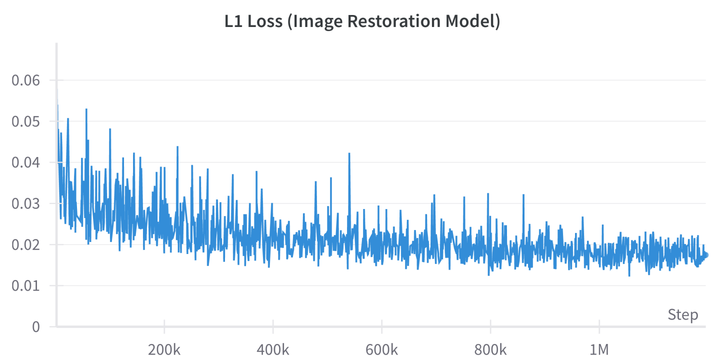


Fig. 1: Image Restoration Loss ($\mathcal{L}_1$) computed between the restored image $\hat{x}$ (model's output) and the reference image x.

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. As we discuss in the paper, adding this task helps the main task of deblurring.

Contemporary Works and Reproducibility. Note that PromptIR, ProRes [18] and Amirnet [31] are contemporary works (presented or published by Dec 2023). We compare mainly with AirNet [14] since the model and results are open-source,

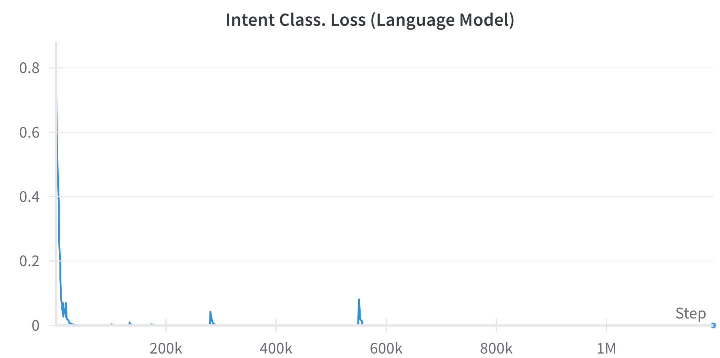


Fig. 2: Intent Classification Loss from the instructions. Product of our simple MLP classification head using **e**. When $\mathcal{L}_{ce} \rightarrow 0$ the model uses the learned prompt embeddings, and it is optimized mainly using the image regression loss ($\mathcal{L}_1$).

and it is a reference all-in-one method. To the best of our knowledge, IDR [33] and ADMS [22] do not provide open-source code, models or results, thus we cannot compare with them qualitatively.

A.1 Additional Ablation Studies

We provide ablation studies and comparison with more task-specific methods in Tables 2 (image denoising) and Table 3 (image deblurring and dehazing).

Table 2: Comparison with general restoration and all-in-one methods (*) at **image denoising**. We report PSNR on benchmark datasets considering different σ noise levels. Table based on [33].

	CBSD68 [19]			**Urban100 [10]**			**Kodak24 [9]**		
Method	**15**	**25**	**50**	**15**	**25**	**50**	**15**	**25**	**50**
IRCNN [35]	33.86	31.16	27.86	33.78	31.20	27.70	34.69	32.18	28.93
FFDNet [36]	33.87	31.21	27.96	33.83	31.40	28.05	34.63	32.13	28.98
DnCNN [34]	33.90	31.24	27.95	32.98	30.81	27.59	34.60	32.14	28.95
NAFNet [4]	33.67	31.02	27.73	33.14	30.64	27.20	34.27	31.80	28.62
HINet [5]	33.72	31.00	27.63	33.49	30.94	27.32	34.38	31.84	28.52
DGUNet [20]	33.85	31.10	27.92	33.67	31.27	27.94	34.56	32.10	28.91
MIRNetV2 [30]	33.66	30.97	27.66	33.30	30.75	27.22	34.29	31.81	28.55
SwinIR [15]	33.31	30.59	27.13	32.79	30.18	26.52	33.89	31.32	27.93
Restormer [29]	34.03	31.49	28.11	33.72	31.26	28.03	34.78	32.37	29.08
* DL [8]	23.16	23.09	22.09	21.10	21.28	20.42	22.63	22.66	21.95
* T.weather [25]	31.16	29.00	26.08	29.64	27.97	26.08	31.67	29.64	26.74
* TAPE [16]	32.86	30.18	26.63	32.19	29.65	25.87	33.24	30.70	27.19
* AirNet [14]	33.49	30.91	27.66	33.16	30.83	27.45	34.14	31.74	28.59
* IDR [33]	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 3: Deblurring and Dehazing comparisons. We compare with task-specific classical methods on benchmark datasets.

Deblurring GoPro [21]		**Dehazing SOTS [13]**	
Method	**PSNR/SSIM**	**Method**	**PSNR/SSIM**
Xu *et al.* [28]	21.00/0.741	DehazeNet [2]	22.46/0.851
DeblurGAN [11]	28.70/0.858	GFN [24]	21.55/0.844
Nah *et al.* [21]	29.08/0.914	GCANet [3]	19.98/0.704
RNN [32]	29.19/0.931	MSBDN [7]	23.36/0.875
DeblurGAN-v2 [12]	29.55/0.934	DuRN [17]	24.47/0.839
InstructIR-**5D**	29.40/0.886	*InstructIR*-**5D**	27.10/0.956
InstructIR-**6D**	**29.73/0.892**	*InstructIR*-**3D**	**30.22/0.959**

B Additional Visual Results

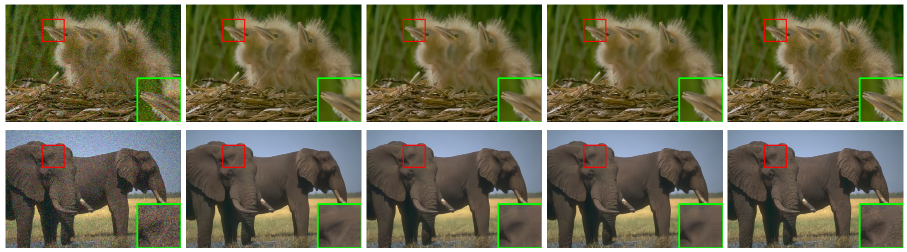
We present diverse qualitative samples in Figures 3, 4. Our method produces high-quality results given images with any of the studied degradations. In most cases the results are better than the reference all-in-one model AirNet [14], and the recent SOTA PromptIR [23]. Also we compare with InstructPix2Pix [1] (diffusion-based) in Figure 6 using real-world cases. In Figure 5, we test our method on real-world samples for image dehazing.

B.1 Efficiency Analysis

We can *process FHD images under 1s* on consumer-grade GPUs (12-24Gb). We are also notably faster and more efficient than the SOTA method PromptIR [23] with 2x less parameters (16M vs. 35M), and 1.6x less operations.

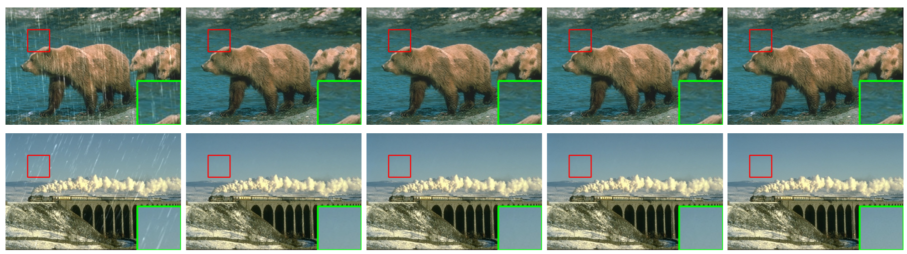
Table 4: Inference cost comparison. Some numbers are from [4].

Method	MPRNet	MIRNet	Restormer	PromptIR	NAFNet	**InstructIR**
MACs(G)	588	786	140	160	65	100



Input AirNet [14] PromptIR [23] *InstructIR* Reference

Fig. 3: Denoising results for all-in-one methods. Images from BSD68 [19] with noise level $\sigma = 25$.



Input AirNet [14] PromptIR [23] *InstructIR* Reference

Fig. 4: Image deraining comparisons for all-in-one methods on images from the Rain100L dataset [8].

Input Hazy RIDCP [27] *InstructIR*

Input Hazy PSD [6] *InstructIR*

Fig. 5: Real Image dehazing comparisons. These are real-world samples without ground-truth. Our method achieves pleasant results as generative models such as RIDCP [27] based on VQGAN. Sample from the RTTS dataset [13]. We use the instruction "remove and haze and mist from this photo please".

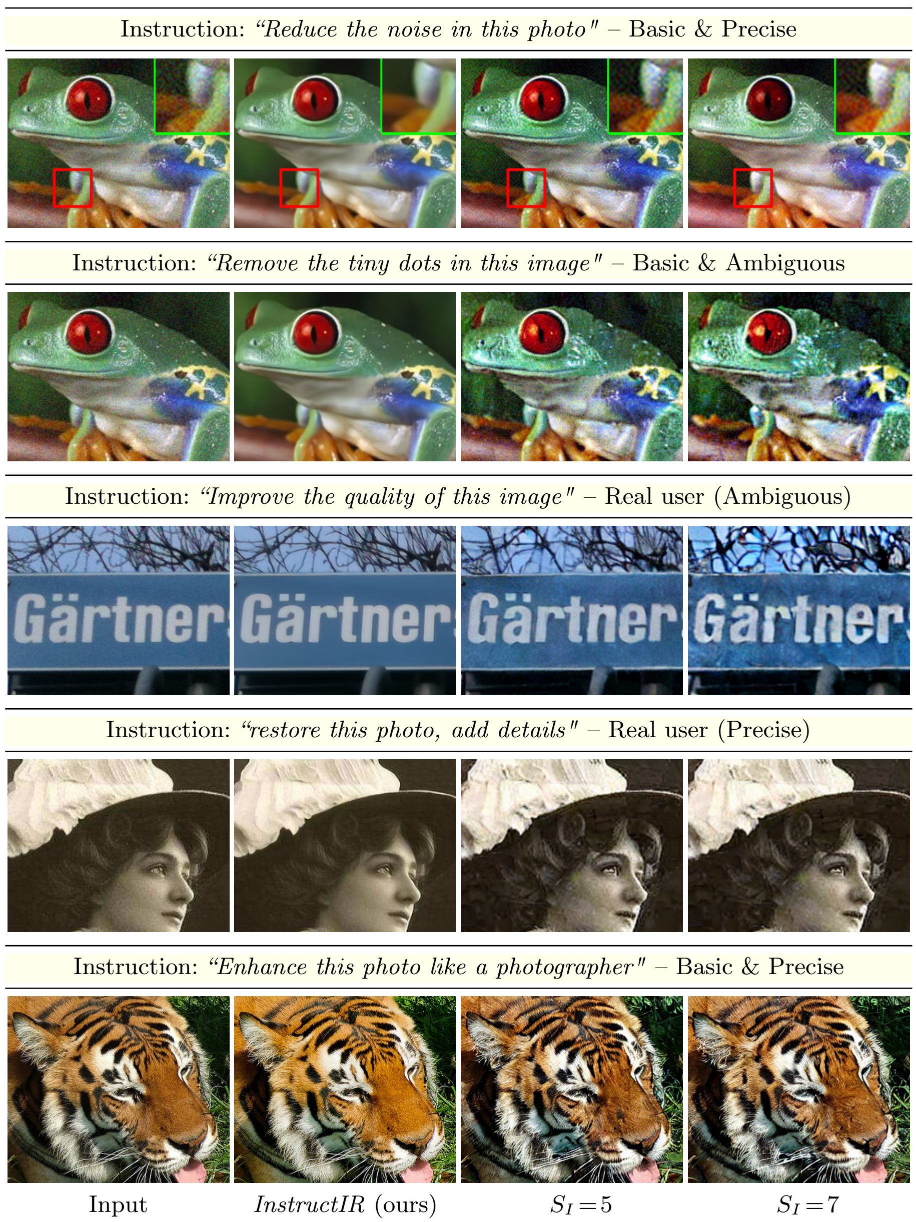


Fig. 6: Comparison with [1] for instruction-based restoration using the prompt. Real-world samples from the *RealSRSet* [15, 26]. We use our **7D** variant. We run [1] using two configurations where we vary the weight of the image component hoping to improve fidelity: $S_I = 5$ and $S_I = 7$ (also known as Image CFG), this parameters helps to enforce fidelity and reduce hallucinations.

References

1. 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 4, 6
2. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing **25**(11), 5187–5198 (2016) 4
3. Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., Hua, G.: Gated context aggregation network for image dehazing and deraining. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1375–1383. IEEE (2019) 4
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV (2022) 2, 3, 4
5. 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) 3
6. Chen, Z., Wang, Y., Yang, Y., Liu, D.: Psd: Principled synthetic-to-real dehazing guided by physical priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7180–7189 (2021) 5
7. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2157–2167 (2020) 4
8. 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) 3, 5
9. Franzen, R.: Kodak lossless true color image suite. http://r0k.us/graphics/kodak/ (1999), online accessed 24 Oct 2021 3
10. 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) 3
11. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: DeblurGAN: Blind motion deblurring using conditional adversarial networks. In: CVPR (2018) 4
12. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In: ICCV (2019) 4
13. 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) 4, 5
14. 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, 4, 5
15. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: ICCV Workshops (2021) 3, 6
16. 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) 3

17. Liu, X., Suganuma, M., Sun, Z., Okatani, T.: 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. pp. 7007–7016 (2019) 4
18. 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
19. 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) 3, 5
20. 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) 3
21. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017) 4
22. 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) 3
23. 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) 4, 5
24. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., Yang, M.H.: Gated fusion network for single image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3253–3261 (2018) 4
25. 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
26. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: enhanced super-resolution generative adversarial networks. In: ECCV Workshops (2018) 6
27. Wu, R.Q., Duan, Z.P., Guo, C.L., Chai, Z., Li, C.: Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22282–22291 (2023) 5
28. Xu, L., Zheng, S., Jia, J.: Unnatural l0 sparse representation for natural image deblurring. In: CVPR (2013) 4
29. 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) 3
30. 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) 3
31. 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) 2
32. Zhang, J., Pan, J., Ren, J., Song, Y., Bao, L., Lau, R.W., Yang, M.H.: Dynamic scene deblurring using spatially variant recurrent neural networks. In: CVPR (2018) 4
33. 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) 3

34. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. TIP (2017) 3
35. Zhang, K., Zuo, W., Gu, S., Zhang, L.: Learning deep CNN denoiser prior for image restoration. In: CVPR (2017) 3
36. Zhang, K., Zuo, W., Zhang, L.: FFDNet: Toward a fast and flexible solution for CNN-based image denoising. TIP (2018) 3