

Supplementary Materials

Image-free Pre-training for Low-Level Vision

Anonymous Author(s)

In the supplementary file, we provide more supporting materials. First, we verify that IFP exhibits robustness to perturbations in the frequency domain, and this capability is data-agnostic. Second, we show the effectiveness of IFP on low-cost task and CNN backbones. Third, we introduce more details on fine-tuning. Finally, more visual results are presented.

1 ROBUSTNESS OF IFP TO SPECTRAL PERTURBATIONS

To validate the ability of IFP to enable the model to extract robust spectral representations, we utilize Restormer [8] as the backbone and conduct IFP pre-training for 30k iterations using a randomly generated Gaussian noise image. Following pre-training, we fix the model parameters and replace the masked noise input with masked real images. The output results are displayed in Fig. 7. From the perspective of the frequency domain, the model now possesses the capability to reconstruct the masked frequency bands. This capability is advantageous for the model to resist spectral perturbations caused by degradation information in downstream tasks, thereby extracting robust representations. From the perspective of the spatial domain, the model now possesses a certain level of image reconstruction ability, and some image shadows caused by spectral damage are mitigated. Since it is unlikely for a single Gaussian noise image to contain any content information related to inputs, and in Sec.4.5 we verify that the effectiveness of IFP does not stem from the content of the image, these experiments provides strong evidence that the favored initialization learned by IFP is data-agnostic. This is consistent with the research conclusions of Kong et al. [3]. Therefore, it is not advisable to try to enhance the performance of IFP by simply increasing the amount of data. This operation also contradicts our original intention of designing pre-training methods for low-level vision tasks in situations of data scarcity.

2 MORE EXPERIMENTAL RESULTS

Low-cost Task. Not only for high-cost tasks, we claim that IFP is also effective for low-cost tasks where degradation is easy to synthesize. We perform denoising experiments with additive white Gaussian noise and the results are shown in Table 2. Following DegAE [4], we train Restormer [8] with noise levels sampled from a wide range of [0, 50] rather than focusing on a specific single noise level for better universality. After fine-tuning, we test the trained models on Kodak24, CBSD68 [5], and Urban100 [1] datasets with different noise levels. Similar to the results in SR, IFP outperforms the state-of-the-art DegAE [4] method on all three datasets.

Effectiveness of IFP on CNN backbones. In Tab. 1, we provide the results of the IFP integrated with CNN backbones on downstream low-light enhancement task, where the IFP paradigm is agnostic to the model architecture.

Table 1: Low-light enhancement results (PSNR) on LOL dataset.

Methods	SSIE [7]	SSIE (IFP)	EnlightenGAN [2]	EnlightenGAN (IFP)
PSNR	18.40	19.21 (+0.81)	20.33	20.43 (+0.10)

Table 2: Image Gaussian denoising results (PSNR).

Method	Kodak24		CBSD68		urban100	
	$\sigma=25$	$\sigma=15$	$\sigma=25$	$\sigma=15$	$\sigma=25$	$\sigma=15$
DnCNN [10]	27.19	31.24	26.10	30.26	25.28	29.81
IRCNN [11]	27.33	31.37	26.25	30.37	25.44	29.93
SRResNet [6]	27.98	32.00	26.72	30.83	26.40	31.02
DRUNet [9]	28.13	32.18	26.48	30.72	26.53	31.17
Restormer	30.13	29.94	28.96	28.73	29.18	28.85
Restormer (DegAE)	30.29	30.14	28.97	28.85	29.23	29.00
Restormer (IFP)	30.38	30.25	29.11	28.97	29.41	29.16
Uformer	30.08	29.97	28.86	28.74	29.13	28.92
Uformer (IFP)	30.20	30.07	28.94	28.80	29.13	28.92

3 MORE DETAILS ON FINE-TUNING

At the downstream task fine-tuning stage, we initialize the model with encoder parameters obtained during the pre-training stage, and replace the decoder with a simple convolutional layer, whose kernel size is 3×3 and the output channel is 3. Following DegAE [4], we set the initial learning rate as $3e-4$ and cosine decayed to $1e-6$. For all downstream tasks, we adopt L1 loss and set the input patch size as 160×160 . The AdamW optimizer with a batch size of 4 is used to train and all downstream tasks are finetuned on a single NVIDIA Geforce RTX 3090 GPU for fair comparison.

4 ADDITIONAL VISUAL RESULTS

We provide more visual results of downstream tasks in Fig. 1 to 6, including low-light image enhancement, image deblurring, and image deraining. It can be observed that our IFP helps remove the low-light/blur/rain more thoroughly and exhibits superior visual recovery quality in all types of degradation.

REFERENCES

- [1] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 5197–5206. <https://api.semanticscholar.org/CorpusID:8282555>
- [2] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2019. EnlightenGAN: Deep Light Enhancement Without Paired Supervision. *IEEE Transactions on Image Processing* 30 (2019), 2340–2349. <https://api.semanticscholar.org/CorpusID:189928152>
- [3] Xiangwen Kong and Xiangyu Zhang. 2022. Understanding Masked Image Modeling via Learning Occlusion Invariant Feature. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 6241–6251.

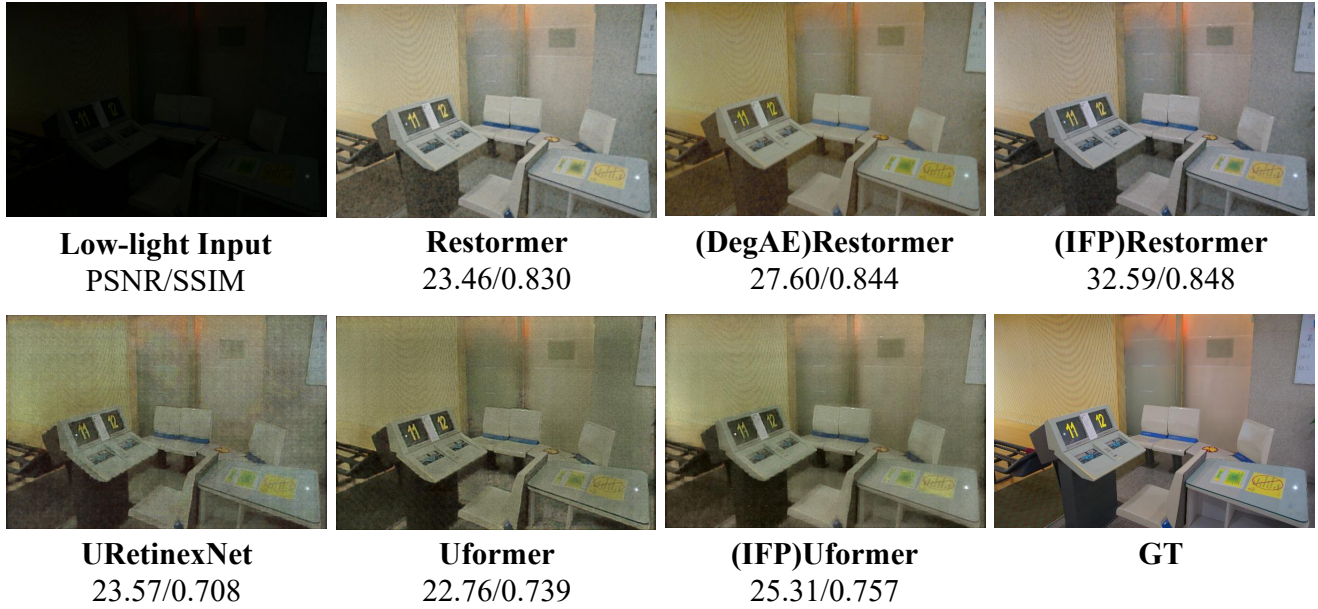


Figure 1: Visual comparison with state-of-the-art methods on LOL dataset. Please zoom in for details.

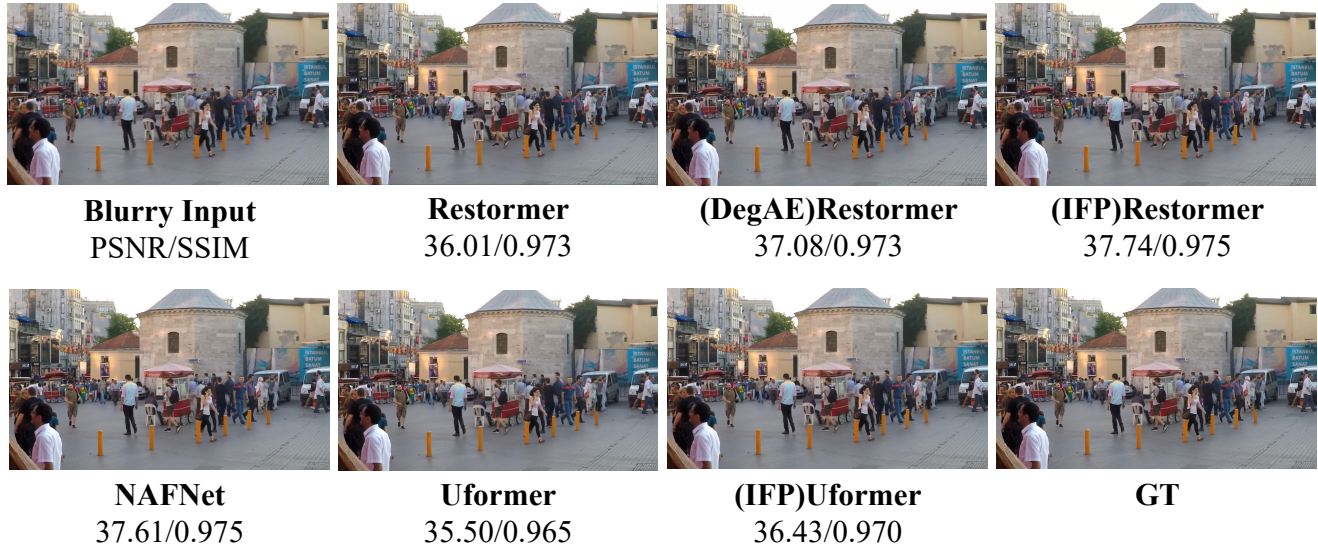


Figure 2: Visual comparison with state-of-the-art methods on Gopro dataset. Please zoom in for details.

- <https://api.semanticscholar.org/CorpusID:251402519>
- [4] Yihao Liu, Jingwen He, Jinjin Gu, Xiangtao Kong, Y. Qiao, and Chao Dong. 2023. DegAE: A New Pretraining Paradigm for Low-Level Vision. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 23292–23303. <https://api.semanticscholar.org/CorpusID:261081381>
- [5] David R. Martin, Charles C. Fowlkes, Doron Tal, and Jitendra Malik. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001 2* (2001), 416–423 vol.2. <https://api.semanticscholar.org/CorpusID:64193>

- [6] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *ECCV Workshops*. <https://api.semanticscholar.org/CorpusID:52154773>
- [7] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep Retinex Decomposition for Low-Light Enhancement. *ArXiv abs/1808.04560* (2018). <https://api.semanticscholar.org/CorpusID:52008443>
- [8] Syed Waqas Zamir, Aditya Arora, Salman Hameed Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2021. Restormer: Efficient Transformer for High-Resolution Image Restoration. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 5718–5729. <https://api.semanticscholar.org/CorpusID:52008443>

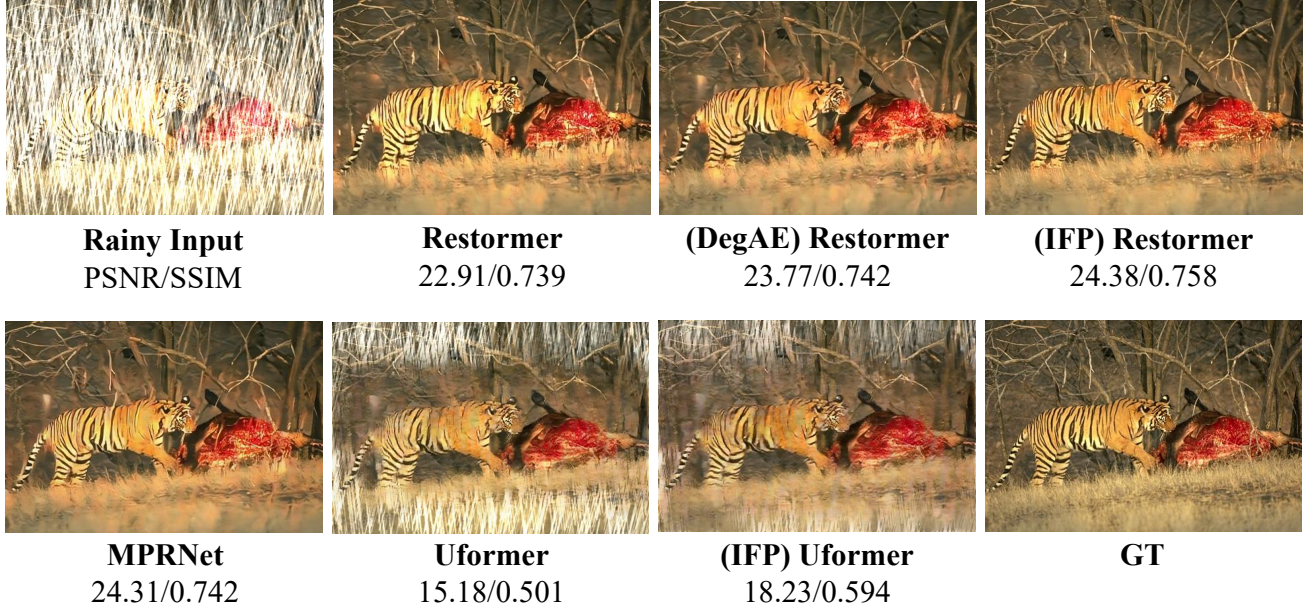


Figure 3: Visual comparison with state-of-the-art methods on Rain100H dataset. Please zoom in for details.

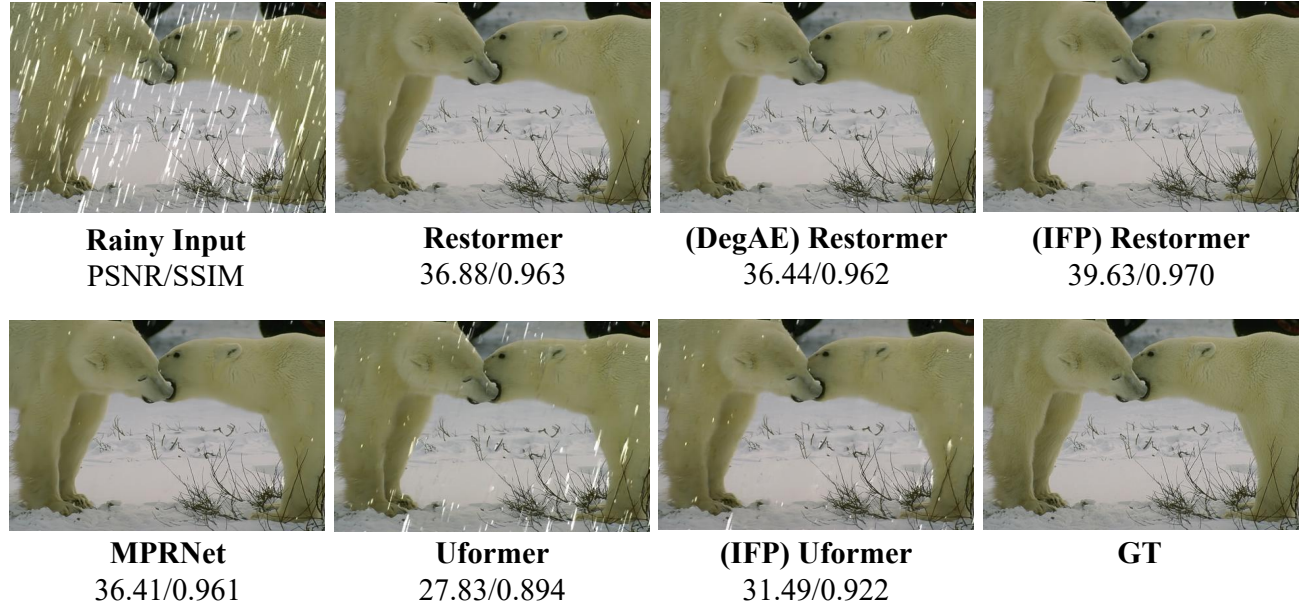


Figure 4: Visual comparison with state-of-the-art methods on Rain100L dataset. Please zoom in for details.

- [9] K. Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. 2020. Plug-and-Play Image Restoration With Deep Denoiser Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2020), 6360–6376. <https://api.semanticscholar.org/CorpusID:221377171>

- [10] K. Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2016. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing* 26 (2016), 3142–3155. <https://api.semanticscholar.org/CorpusID:996788>
- [11] K. Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. 2017. Learning Deep CNN Denoiser Prior for Image Restoration. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 2808–2817. <https://api.semanticscholar.org/CorpusID:996788>

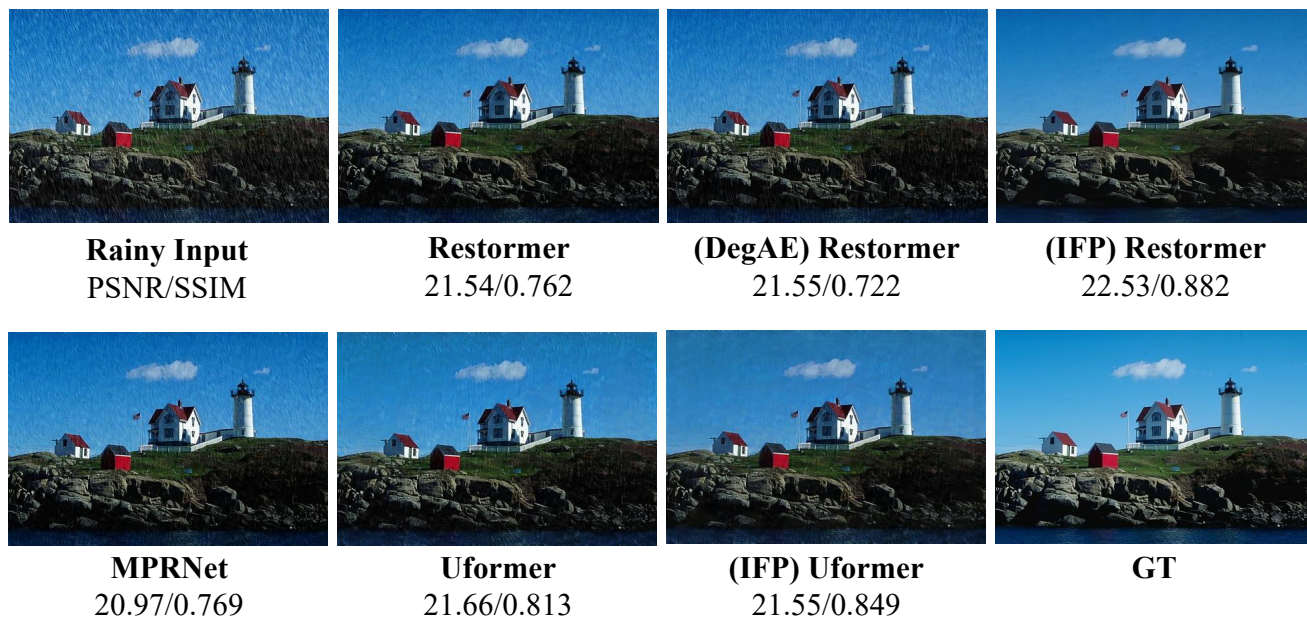


Figure 5: Visual comparison with state-of-the-art methods on Test100 dataset. Please zoom in for details.

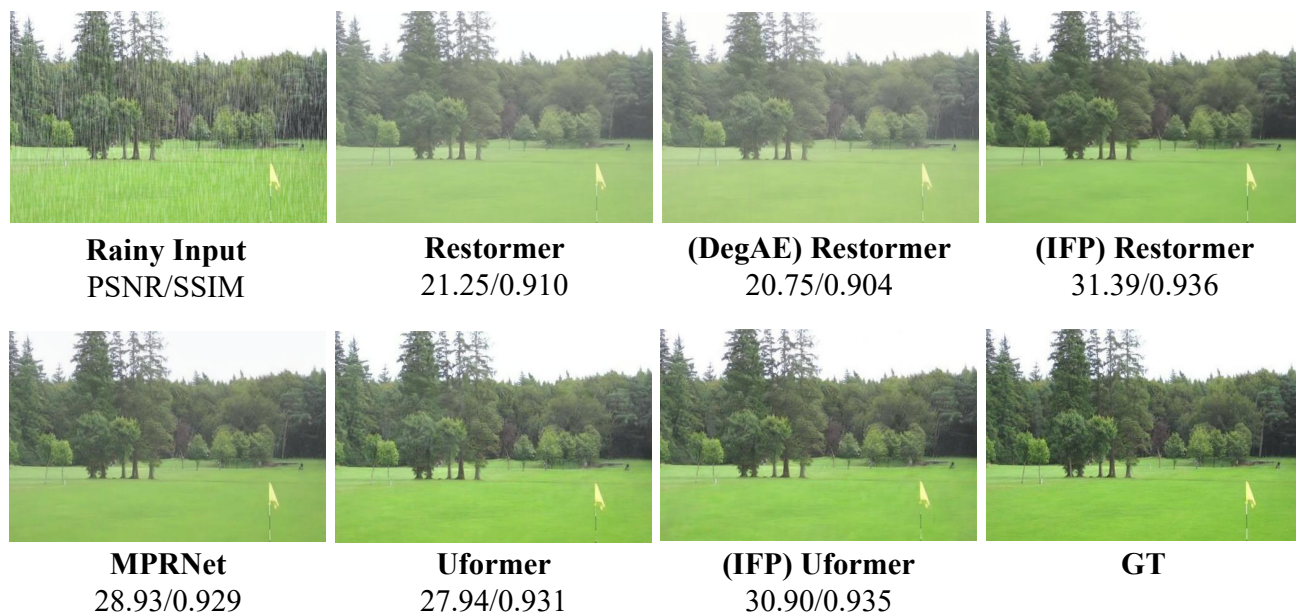


Figure 6: Visual comparison with state-of-the-art methods on Test1200 dataset. Please zoom in for details.

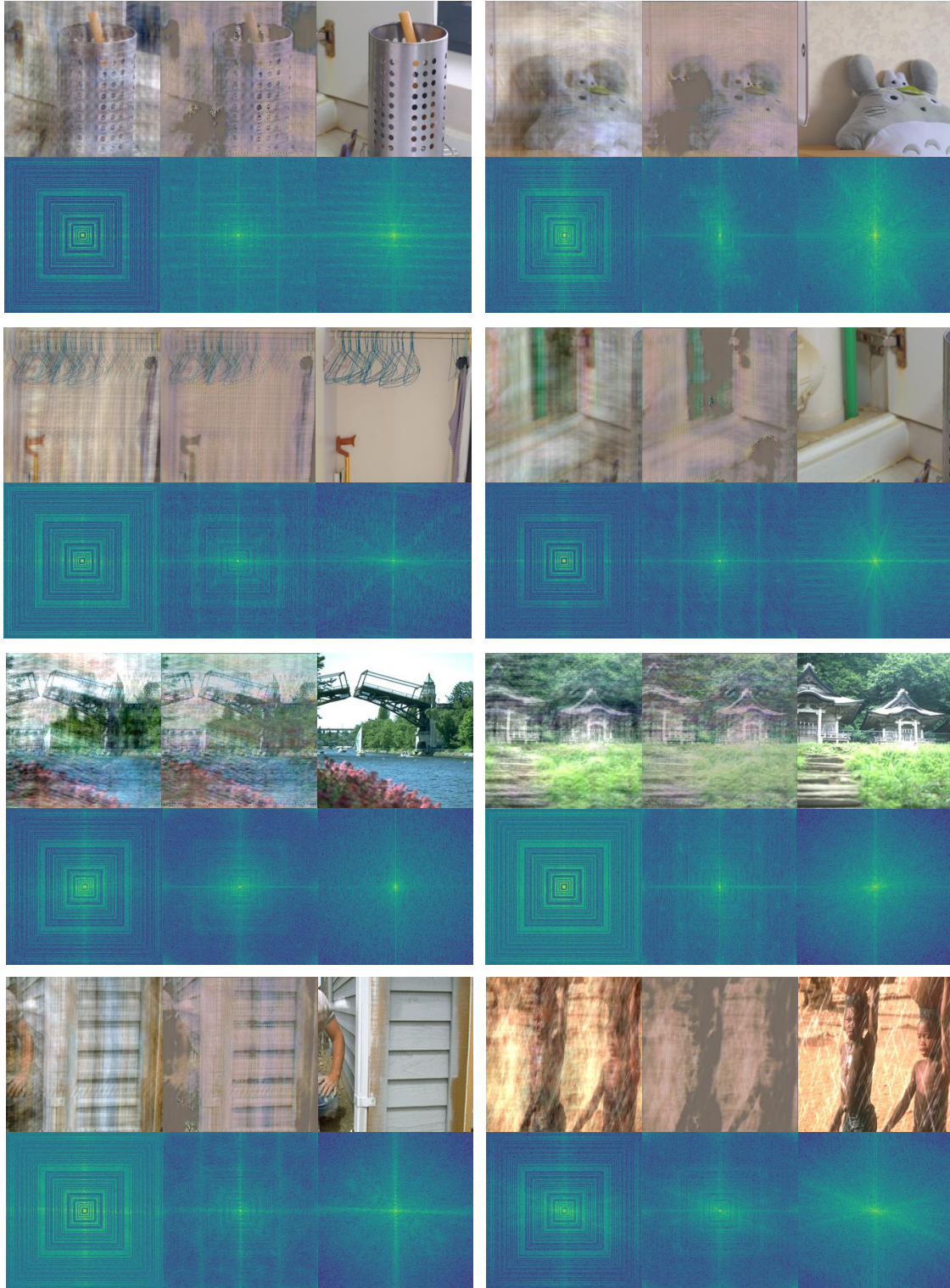


Figure 7: Reconstructions using IFP pretrained with a randomly generated Gaussian noise image. For each sextuplet, we display the spatial results (top) and frequency results (bottom). For each triplet in a specific domain, we display the masked input (left), our IFP reconstruction (middle), and the ground truth (right).