UNICORN: Latent Diffusion-based Unified Controllable Image Restoration Network across Multiple Degradations

Supplementary Material

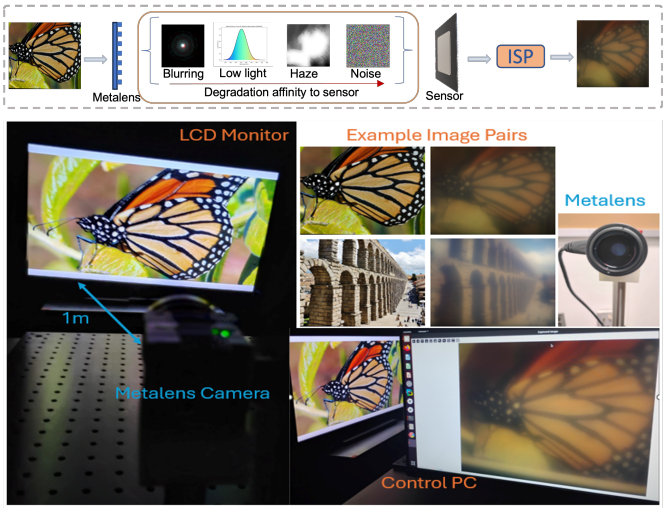


Figure 1. METARESTORE data capture setup

1. Metalens Image Formation

For a given source image S, the imaging process for a camera with a wavelength-dependent point spread function (PSF) p_λ can be described as:

$$I_c(x,y) = \mathcal{P}\Big(\int (\rho_\lambda(S) * p_\lambda)(x,y)\kappa_c(\lambda)\, d\lambda \Big) + \mathcal{N} \quad (1)$$

where $I_c \in \mathbb{R}^{H\times W\times C}$ represents the sensor measured image with c denoting the RGB channels, $\rho_\lambda(\cdot) < 1$ is the lens's relative light efficiency, $\kappa_c(\lambda)$ the spectral sensitivity of channels to wavelength λ, and $*$ denotes convolution. $\mathcal{P}(\cdot)$ and $\mathcal{N}$ represents Poisson noise and additive Gaussian noise, respectively. The source image (S) can then be estimated as $\hat{S} = \arg\max_S(S|I) = \arg\max_S(I|S)(S)$ using a *maximum a-posteriori* framework with a known prior (S).

Restoration of image S from sensor measurement I introduces severe artifacts due to the ill-posed nature of the image formation process (Eq. (1)), especially impacted by the sensor's spectral response $\kappa_c(\lambda)$ and noise $\mathcal{N}$. Ignoring $\kappa_c(\lambda)$ reduces the image restoration problem to deblurring with a known PSF, but integrating wavelength-dependent κ_c over λ introduces haze in the image, requiring both deblurring and dehazing. Sensor noise further degrades high frequencies in the image, leading to loss of fine details. Therefore, to simplify the restoration problem, we pose Eq. (1) as

$$I(x,y) = \mathcal{H}(\mathcal{B}(\mathcal{D}(S))) + \eta \quad (2)$$

where $\mathcal{H}(\cdot)$, $\mathcal{B}(\cdot)$, $\mathcal{D}(\cdot)$ and η represent hazing, blurring, image darkening and the noise. Note that reconstructing S from simplified Eq. (2) still remains a challenging non-linear problem, beyond the scope of traditional deblurring, denoising, dehazing or contrast enhancement techniques.

2. Implementation details

2.1. Synthetic single datasets

For obtaining an all-in-one restoration model for multiple degradations, as proposed in the main manuscript, we augment our training with synthetic single degradation datasets like blurring with a point spread function like Gaussian (with different standard deviations) or Jinc which is quite different from the motion blur encountered in GoPro [9]. Similarly, we differentiate severe low-light enhancement for images with zero visibility as in Sony Total Dark [15] against moderate low light (or contrast) enhancement from LOL [14]. We present qualitative comparisons for such special but frequently observed degradations types in Fig. 10 and Fig. 11.

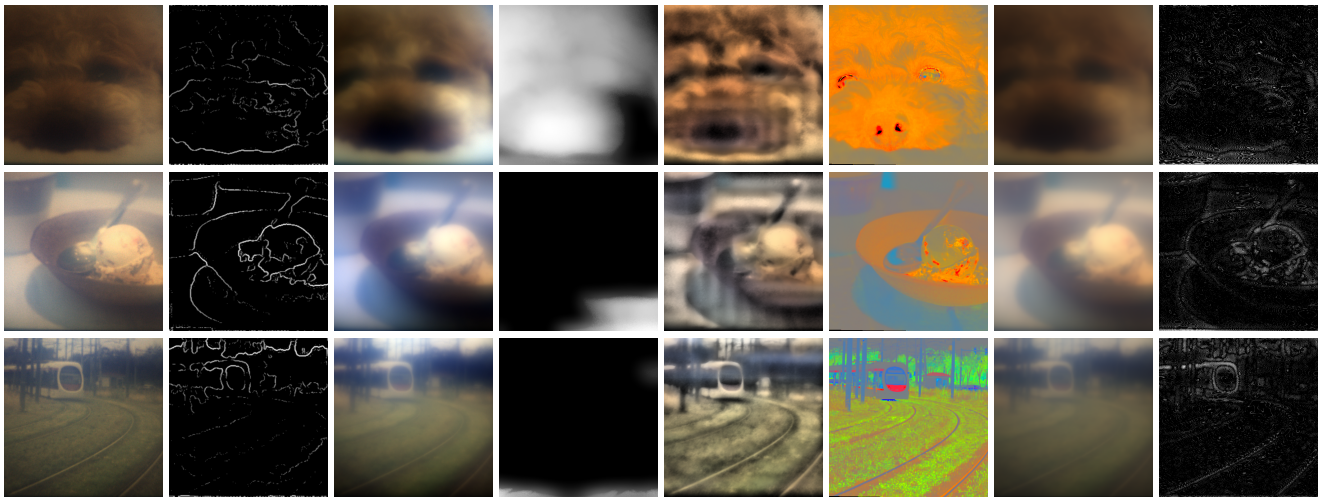
2.2. Synthetic mixed datasets

As mentioned in the Sec. 4 of the main manuscript, we evaluated our method on three synthetic mixed datasets and summarized the results in Tab. 2 of the main paper. For a fair comparison against state of the art methods, we evaluate our model trained on single degradation datasets for deblurring, dehazing, denoising and low light enhancement followed by a short fine-tuning phase on a small subset ($< 10\%$) of the full mixed degradation dataset. We compare against the publicly released checkpoints of DiffUIR(-L) [17] for its full all-in-one model, AutoDIR [2] and DA-CLIP [7] for its in-the-wild model. Figure 12, Fig. 13 and Fig. 14 present our qualitative results for the mixed degradation restoration.

2.3. METARESTORE dataset creation

Metalens Fabrication We fabricated the 1 cm large aperture meta-optic using a nanofabrication approach, facilitated in an ISO Class 5-7 clean room environment. First, quartz wafer (purchases from University Wafer), were cleaned in subsequent ultrasonicating baths of Acetone, IPA, and DI water, then exposed to a short oxygen descum in a Barrel Etcher. We then deposited an $\sim$ 800 nm thick Silicon Nitride film using plasma enhanced chemical vapor deposition in an SPTS chamber. Afterwards the wafer was

Figure 2. Low level guidances for metalens images



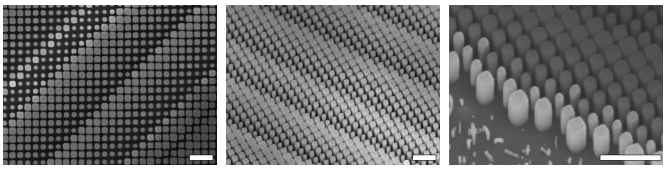


Figure 3. Metalens SEM images

diced into 1.5 cm square pieces and again cleaned using an ultrasonicating bath and barrel etch steps (as before). We then applied a positive resist (ZEP 520A) with a thickness of 400nm. To mitigate charging during the patterning, we also applied a conductive polymer layer (DisCharge H2O). We then patterned the resist using a 8 nA, 100 keV electron beam (JEOL JBX6300FS) at a dose of about 300 µC cm^2. After electron beam lithography, we removed the conductive polymer layer using a short IPA bath and developed the resist at room temperature in Amyl Acetate for 2 min. Subsequently, the sample was again descummed in a short barrel etch step and a layer of about 75 nm alumina was evaporated onto the sample. The resist was then lifted off overnight in an NMP bath at on a hot plate. Subsequently, the SiN layer was etched using a fluorine based etch mixture in an inductively coupled reactive ion etcher (Oxford PlasmaLab System 100). Finally, the chip was integrated in a 3D printed holder and mounted with the sensor. Scanning Electron Microscope Images after fabrication can be seen in Fig. 3.

Data Capture A 405nm metalens camera was placed 1 m from an LCD display to maintain consistent optical path length, after calibrating the exposure time, white balance, and contrast to optimize image quality. The process was conducted in a dark room to eliminate ambient light interference. This is to make sure that the captured images reflected only the distortions introduced by the metalens. Using the Div2K [1] dataset, 800 images (training dataset) were displayed on the screen and captured by the camera, with a Linux server and the Vimba SDK automating frame alignment, timing, and camera control for a consistent acquisition. Similarly, we captured new 400 images from [6] as the evaluation dataset for our method. The captured images show a mixture of degradation types, including chromatic aberrations and blur, as shown in figures displayed. The overall setup is presented in Fig. 1.

2.4. Low level cue generation

Section 3.2 explains the process of generation of several primary I_0^h and secondary $I_{1,2,...}^h$ guidances to refine the overall control. Figure 2 visualizes all the different cues we generate for the mixed degradation encountered in metalenses. While each primary guidance itself does not look close to the original image, their effect on the UNICORN control head coupled with low level information from the secondary guidances helps image restoration in both seen (Fig. 5) as well as unseen scenarios (Fig. 4).

2.5. Model details

Training and inference settings We initialize ControlNet with Stable Diffusion (SD) backend after switching the prompt embedder to a image-to-image CLIP model instead of text-to-image. The fine-tuned image-to-image SD checkpoints are obtained from the Diffusers library [13]. This forms our baseline model.

For inference, we use the DDIM [12] sampling schedule over 50 iterations. Our proposed model runs on a single

consumer NVIDIA RTX 3090 24GB GPU with an average inference time of 7 seconds across all tasks. At the end of the reverse diffusion process, we perform an optional histogram matching with the source image (or one of the primary guidances depending on which appears better) to generate color consistent samples.

2.6. Discussion on MoE and TSU units

To assess the importance of the shared multi-control head architecture we first ablate the K control heads from UNICORN(Fig. 3 main paper), and then replace it with a single control head with low-level input cues similar to ControlNet [16]. For ablating the task stabilizer units (TSU), we simply replace them with identity functions. To assess the contributions of the mixture-of-experts (MoE) adaptors, we replace weighted fusion by naive addition of controls from all control paths.

Quantitative metrics (PSNR) on model performance after ablations.

Datasets	ControlNet [16]	UNICORN	w/o MoE	w/o TSU
Single Tasks	24.09	29.20	29.20	28.14
Mixed Tasks	22.54	28.61	26.33	25.41
METARESTORE	21.36	27.93	23.42	24.10

As shown in the table above, vanilla ControlNet [16] performs poorly compared to our method in the absence of task-specific control pathways due to the sharing of a single control head among multiple degradations. At the same time, naively adding controls in the absence of the MoE units fails to capture the importance of task-specific restoration cues, leading to a drop in the metrics. Finally, removing the TSU units effectively eliminates task sharing among different degradations, making performance on mixed degradations restoration brittle.

2.7. Choice of low-level cues

The motivation for using low-level cues stems from trying to unlock the power of intermediate image representations from traditional image processing techniques, when used with deep learning techniques. In our UNICORN architecture, we make the choice of augmenting the degraded image with a variety of low-level cues for every degradation type, rather than associating low-level cues with a specific degradation.

2.8. Effect of task-based text prompt

Our base diffusion model is a finetuned image-text mixer variant [13] over the original Stable Diffusion 1.4 model. Although it preserves faint text based task control due to our training scheme involving task-based text prompts, we observe the visual performance to be much better when using a generic task prompt and letting the MoE adaptors decide on the restoration task weighting.

2.9. Evaluation Fairness Assessment

Our experiments use open-source code and models from baseline methods. The performance drop in baseline methods from Table 4 (single tasks) to Table 1 (zero-shot) is due to the limitations of existing methods in handling multiple simultaneous degradations. This trend is also supported by the experimental observations in DiffUIR[1] [17], where AirNet [4], PromptIR [10], and DA-CLIP [7] perform poorly on real-world datasets. AutoDIR's sequential restoration is sub-optimal as seen in Fig. 2 in the main paper. DA-CLIP used a fine-tuned CLIP [11] model to generate preconditioning information on single degradation tasks, which limits it's generalization capability on mixed restoration tasks, as was also noted as a limitation of their work. For AirNet and PromptIR, we use open-source checkpoints for denoising and dehazing (Tables 1, 2 and 4) and fine-tune them for other tasks. For AutoDIR, DA-CLIP, and DiffUIR (Tables 1, 2 and 4), we use their all-in-one universal checkpoints. For AutoDIR, we report performance without including their structural consistency model since that operates as an independent degradation restoration module. Fine-tuning AirNet on the METARESTORE benchmark showed no improvements (Table 1).

3. More visualisations

We provide additional visual results for single degradation removal similar to Fig. 4 from the main text for all 4 benchmarks presented in Tab. 1 of the main manuscript.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017. 2

[2] Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. *arXiv preprint arXiv:2310.10123*, 2023. 1, 4, 5

[3] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE TIP*, 2019. 5

[4] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, 2022. 3, 4

[5] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, 2020. 5

[6] J. Liu, D. Liu, W. Yang, S. Xia, X. Zhang, and Y. Dai. A comprehensive benchmark for single image compression artifact reduction. *IEEE TIP*, 2020. 2

[7] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. In *ICLR*, 2024. 1, 3, 4, 5

[1] Refer to Table 4 on unknown degradation restoration.

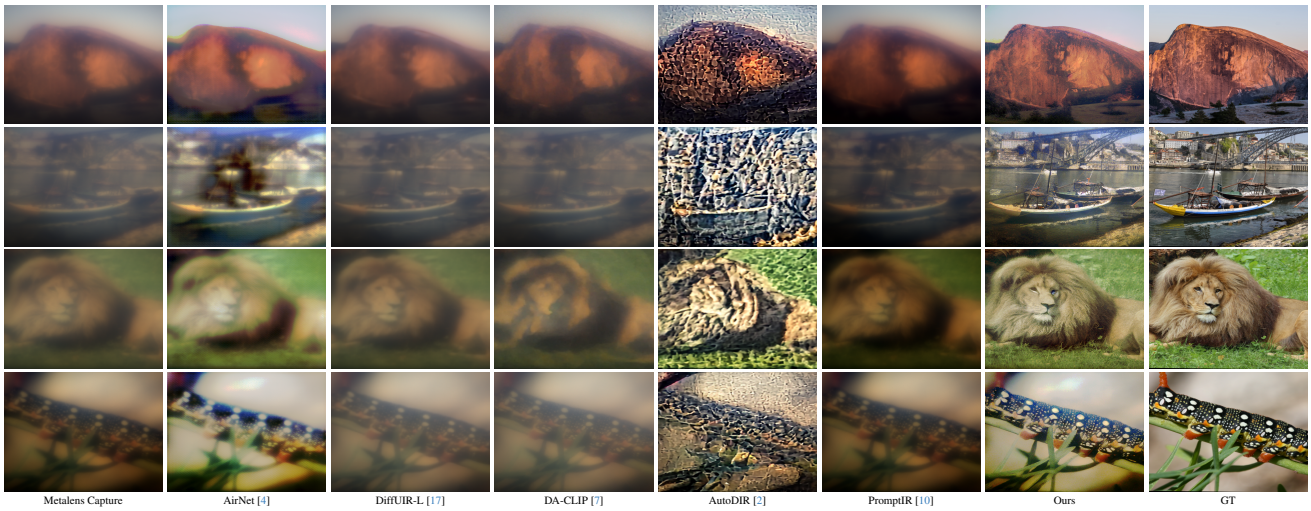


Figure 4. Qualitative results on the proposed METARESTORE metalens imaging benchmark. The results shown here are ***zero-shot***; none of the models have been fine-tuned on the dataset. Best viewed if zoomed in.

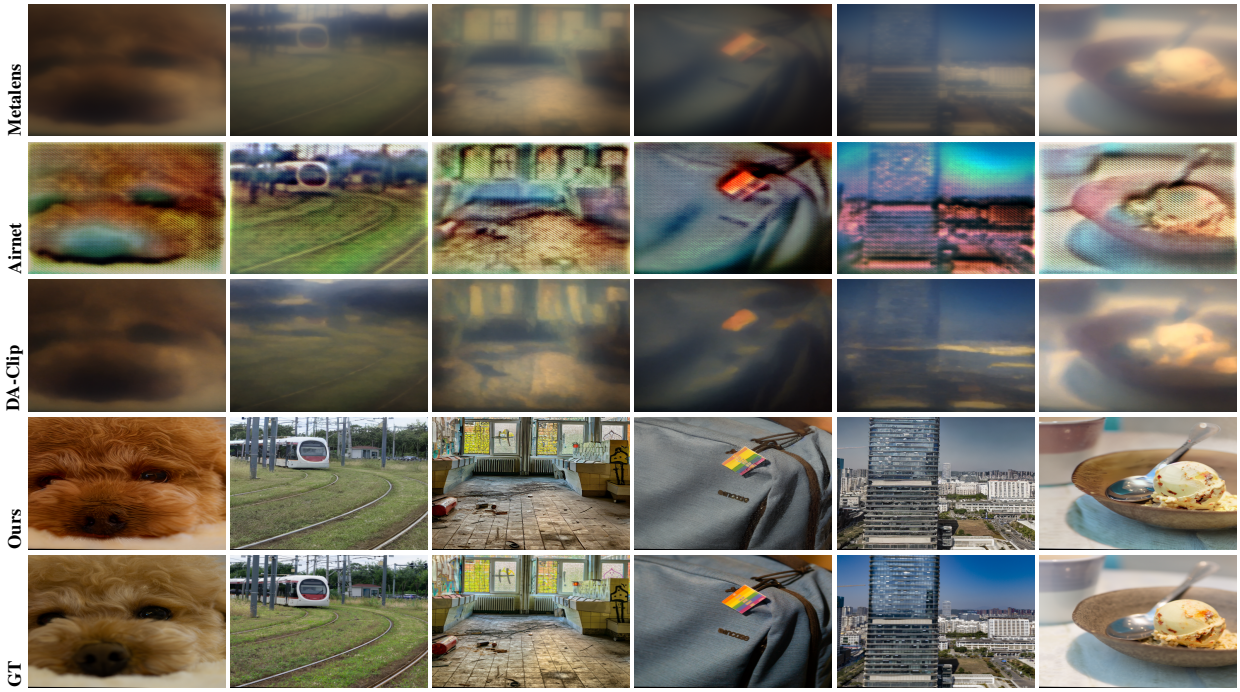


Figure 5. Visual results of models ***fine-tuned*** on METARESTORE.

[8] 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

[9] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 5

[10] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. *NeurIPS*, 2024. 3, 4, 5

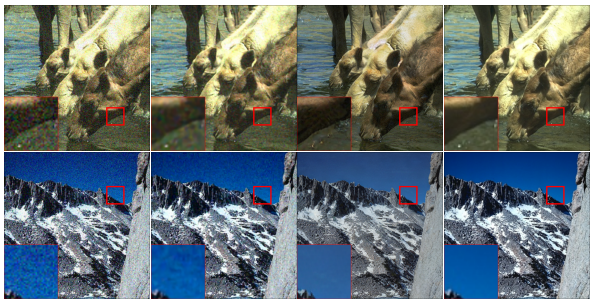
[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[13] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 2, 3

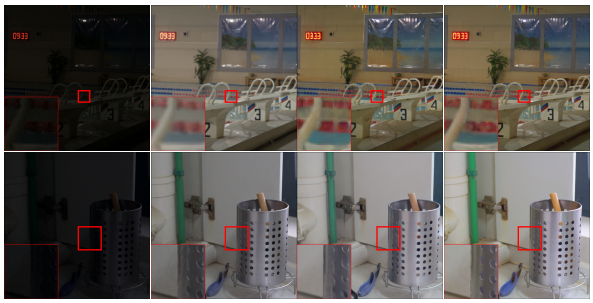
Input AutoDIR [2] Ours Ground truth

Figure 6. Deblurring results on GoPro [9]



Input Airnet [5] Ours Ground truth

Figure 7. Denoising results on CBSD68 [8]



Input DA-CLIP [7] Ours Ground truth

Figure 8. Low light enhancement results for LOL [14]

Input PromptIR [10] Ours (a) Ground truth

Figure 9. Dehazing results for RESIDE [3]

Input DA-CLIP [7] Ours Ground truth

Figure 10. *Extreme* low light enhancement results for SID [15]

Input AutoDIR [2] Ours Ground truth

Figure 11. Non-motion deblurring results

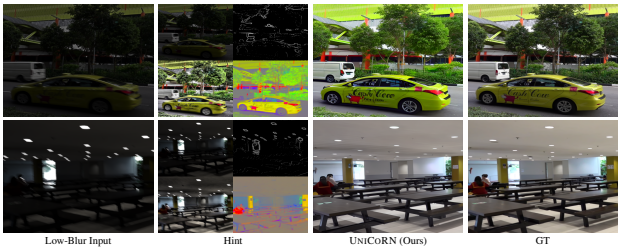


Figure 12. Qualitative results on the low-blur task. Best viewed if zoomed in.

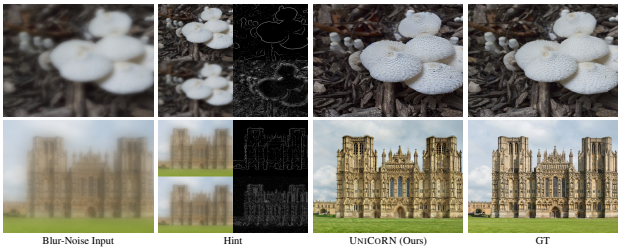


Figure 13. Qualitative results on the blur-noise task. Best viewed if zoomed in.

[14] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 1, 5

[15] Qingsen Yan, Yixu Feng, Cheng Zhang, Pei Wang, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. You only need one color space: An efficient network for low-light image enhancement. *arXiv preprint arXiv:2402.05809*, 2024. 1, 5

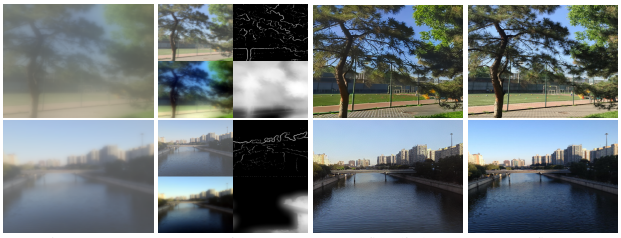


Figure 14. Qualitative results on the blur-haze task. Best viewed if zoomed in.

[16] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[17] Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *CVPR*, 2024. 1, 3, 4