Removing Diffraction Image Artifacts in Under-Display Camera via Dynamic Skip Connection Network Suppplementary Material

Ruicheng Feng[1] Chongyi Li[1] Huaijin Chen[2] Shuai Li[2] Chen Change Loy[1] Jinwei Gu[2,3]
[1]S-Lab, Nanyang Technological University [2]Tetras.AI [3]Shanghai AI Laboratory
{ruicheng002, chongyi.li, ccloy}@ntu.edu.sg
{huaijin.chen, shuailizju}@gmail.com gujinwei@tetras.ai

In this Supplementary Material, we present additional details and discussions for the image formation model and DISCNet proposed in the main body as follows:

- Light Propagation Model
- Incomplete Degradation in LDR Scenes
- Comparison with Previous Dataset
- Training Details
- Limitations
- Visual Results on Simulated Dataset
- Visual Results on Real Dataset

1. Light Propagation Model

The light propagation model in the UDC system can be divided into following steps:

Propagation between the point source and the OLED display. The light emitted from the point light source first hit on the front plane of the OLED display, where the optical field $U_{D-}(p,q)$ can be expressed as

$$U_{D-}(p,q) = \exp\left(\frac{i\pi}{\lambda z_1}(p^2+q^2)\right), \tag{1}$$

where (p,q) is the 2D spatial coordinates, λ is the wavelength and z_1 is the distance between the point light source and the OLED display. We assume that the point source has unit amplitude.

Modulation by the OLED display. The light hit on the front plane of the OLED display will be modulated by its transmission function $t(p,q)$, which is determined by the specific design of the display pattern. The optical field after modulation $U_{D+}(p,q)$ becomes

$$U_{D+}(p,q) = U_{D-}(p,q)t(p,q). \tag{2}$$

Propagation between the OLED display and the lens. The light modulated by the OLED display propagates for a distance of d, before hitting on the front plane of the lens, where the optical field $U_{L-}(p,q)$ can be computed using Fresnel propagation as

$$U_{L-}(p,q) = U_{D+}(p,q) * \exp\left(\frac{i\pi}{\lambda d}(p^2+q^2)\right). \tag{3}$$

Here, $*$ denotes the 2-D convolution operator.

Modulation by the lens. The light hit on the front plane of the camera lens will be modulated by the lens transmission function, which is determined by focal length f of the lens. The optical field after modulation $U_{L+}(p,q)$ becomes

$$U_{L+}(p,q) = U_{L-}(p,q)\exp\left(\frac{-i\pi}{\lambda f}(p^2+q^2)\right). \tag{4}$$

Propagation between the lens and the sensor. The light modulated by the lens propagates for a distance of z_2, before hitting on the sensor, where the optical field $U_S(p,q)$ can be computed using Fresnel propagation as

$$U_S(p,q) = U_{L+}(p,q) * \exp\left(\frac{i\pi}{\lambda z_2}(p^2+q^2)\right). \tag{5}$$

Finally, the PSF of the imaging system is given by

$$k = |U_S|^2. \tag{6}$$

With the above equations, we can theoretically simulate the PSF of a UDC system, given the exact pixel layout of a display. Due to proprietary reasons, we do not have access to the detailed pixel structures of the particular UDC device (ZTE Axon 20) that we used in the main paper. To validate the above light propagation model, we place another commercial OLED display with known pixel layout in front of a normal camera to construct a UDC system, and use it to measure the PSF. In Figure 1, we found that although the simulated and real-measured PSF share a similar shape, they slightly differ in color and contrast due to model approximations and manufacturing imperfections.

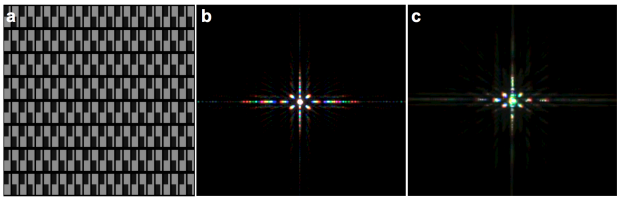


Figure 1. **Comparison of simulated and real-measured PSF.** (a) The pixel layout of a commercial OLED display. Here, different gray-scale values represent different light transmittance of the display. (b) Simulated PSF. (c) Real-measured PSF. The PSFs are brightened to visualize the structured sidelobe patterns.

2. Incomplete Degradation in LDR Scenes

As described in Section 3.2 of the main body, images captured by UDC systems in real HDR scenes will exhibit structured flares near strong light sources. Since the PSF of UDC has a strong response at the center but vastly lower energy at long-tail sidelobes, only when convolved with sufficiently high-intensity scenes, these spike-shaped sidelobes can be amplified to be visible in the degraded image.

Therefore, in an MCIS system proposed in [5], where scenes are displayed on a LCD monitor, which commonly has limited dynamic range, the degradation of a UDC imaging system is incomplete compared to the capture in real HDR scenes. As shown in Figure 2, if the real HDR scene is directly captured with a UDC device, we can observe flare effects near strong light sources. However, for the same scene and same imaging device, the flares are no longer visible in the acquired image if it is displayed on a LCD monitor, since the scene in this case only involves limited dynamic range.

Apart from MCIS data, we also illustrate that HDR scenes are indispensable for our data simulation pipeline. Specifically, if we clip the scene from HDR to LDR, the flare artifacts caused by diffraction effects become invisible in the degraded images (see Figure 3). This further illustrates the importance of HDR scenes. Hence, in order to correctly model the real degradation of a typical UDC system, we involve real HDR scenes in the image formation model in main paper.

3. Comparison with Previous Dataset

In this section, we compare the datasets used in our work with previous one [5] in Table 1. The proposed image formation model could simulate more complex and realistic degradation compared to the dataset in [5].

4. Training Details

Data Simulation for Training. We use the image formation model in main paper to simulate degraded images exhibiting diffraction artifacts. In particular, we set $x_{max} = 500$ for the clipping operation $C(\cdot)$.

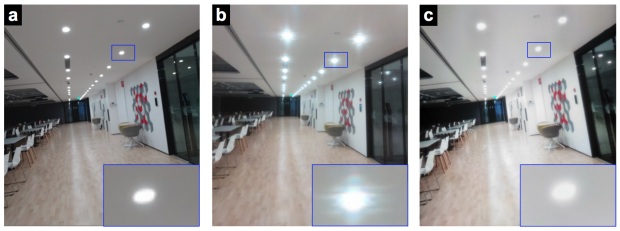
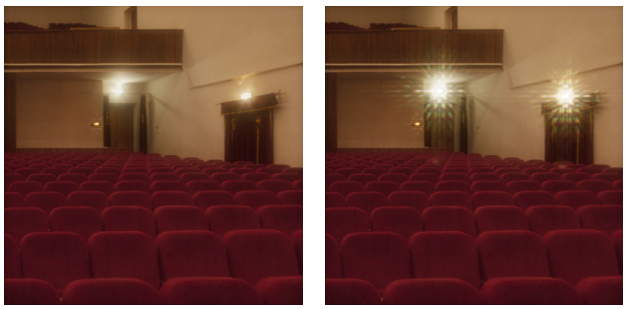


Figure 2. **Comparison of UDC images of real HDR scene and monitor-generated LDR scene.** (a) Real HDR scene captured by a normal camera. (b) Real HDR scene captured by the UDC device. (c) Monitor-generated LDR scene, *i.e.* display the image (a) on a LCD monitor, captured by the UDC device.



(a) Simulated w/ LDR (b) Simulated w/ HDR

Figure 3. Comparison of images simulated with LDR and HDR.

For the tone mapping function, we apply a simple rule [2], given by

$$\phi(x) = \frac{x}{x+\alpha}, \tag{7}$$

where α controls the scale of high luminances. The hyperparameter is set to 0.25 in our case. This formulation mainly compress the high intensities, scaled by approximately $1/x$, and is guaranteed to bring all intensities within displayable range. Many scenes are predominated by a normal dynamic range, but have a few high luminance regions nearby highlights, *e.g.*, street lamp, sunlight. Besides, the sidelobes of the PSF have far lower energy compared to the main peak, leading to relatively low-intensity flare and haze effects in the degraded images. Therefore, this formulation can compress saturated highlights while preserving details in lowlight regions, providing a better display of diffraction artifacts. For simplicity, we mainly focus on analyzing the diffraction effects of UDC and set $n = 0$, providing a noise-free version of the simulated data.

For testing on simulated datasets, we build a test kernel set for quantitative evaluations of different methods. It consists 9 selected rotation variations that are performed on the PSF, *i.e.*, $\{-12, 9, 6, 3, 0, 3, 6, 9, 12\}$. The PSFs first rotate by an angle selected from the above set, and then are convolved with the ground-truth images using image formation model in main paper to generate the corresponding degrade

Table 1. Comparison of different datasets.

Dataset	Zhou *et al*. [5]	Simulated data (Ours)	Real data (Ours)
Scene	Displayed on a monitor	HDR images	Real scenes
Dynamic Range	Low	High	High
Data Format	16-bit RAW	32-bit RGB	14-bit RGB
Major Degradation	Low-light, Color shift	Flare, Haze, Blur, Saturation	Flare, Haze, Blur, Saturation, Veiling glare
UDC System	Lab prototype	-	Commodity UDC production

images. In total each ground-truth image has 9 degraded counterparts, yielding 9 testing sets. Note that only the PSF at the center of the sensor is measured, and the rest in the kernel are generated by applying rotation transformations to the center one. Although it only considers simple variations (rotation) on the PSF, it can still be used to evaluate the performance of non-blind image restoration approaches.

Loss Function. To train the proposed model, we adopt two widely-used losses of image restoration tasks. We originally experimented with $\mathcal{L}_1$ loss between the reconstructed and ground-truth images in tone-mapped domain. To encourage more realistic results, we further apply the perceptual loss in [1], which is defined using the pre-trained VGG-19 network [3] and given by

$$\mathcal{L}_{VGG} = ||\Phi_l(\tilde{x}) - \Phi_l(\hat{x})||_2^2, \tag{8}$$

where Φ_l is the feature maps extracted from the l-th layer of the pre-trained VGG network, and $\tilde{x}$ is the reconstructed image of our network. In particular, we use the "conv5-4" layer as [4]. The total loss for training is formulated by

$$\mathcal{L}_{total} = \mathcal{L}_1 + \lambda \mathcal{L}_{VGG}, \tag{9}$$

where the weight λ is set to 0.01 for balancing the scale of different losses in our experiments.

5. Limitations

Kernel Mismatch. Under the non-blind setting, our method assumes the awareness of the kernel PSF. In real scenarios, however, an estimation of the PSF can be easily affected by noises and artifacts, which results in a kernel mismatch and severely deteriorates the performance of the network. Figure 4 shows the sensitivity of the PSNR performance to the kernel mismatch. In the upper-right and lower-left regions, where the kernels used for simulation and condition differ the most, we can observe huge gaps (over 3 dB) on the PSNR performance. In contrast, the results on the diagonal, where the kernels match precisely, show the best performance in their corresponding rows or columns.

Large and Strong Highlights. While achieving rather satisfactory results on small area of light sources, DISCNet still struggles when highlight regions are large and intensities are extremely strong, leading to over-corrected results. We also conduct experiments on scenes with larger and strong highlights to illustrate the limitations of our method, which is shown in Figure 5. The degraded image contains extremely strong light sources which causes diffraction artifacts in large neighbouring low-intensity regions. Nevertheless, our method is still able to suppress flare and haze

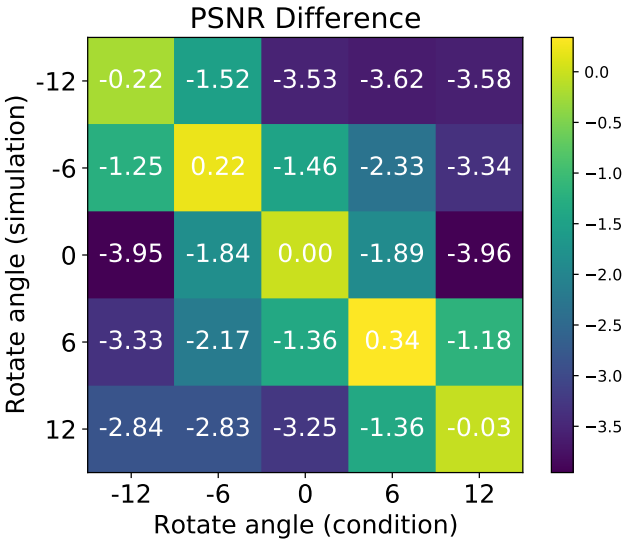


Figure 4. **Recovery sensitivity to the kernel mismatch.** Simulation angles indicate rotations of PSF used in simulation, while condition angles represent the ones used as conditions to DISCNet.

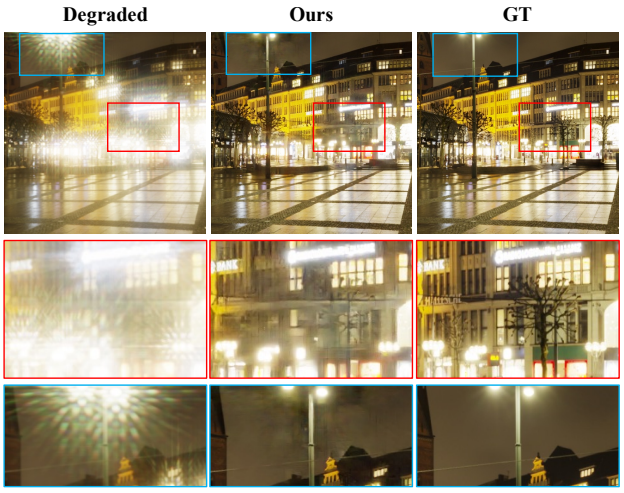


Figure 5. Failure cases around strong light sources.

effects and recover lost details in most regions, even when there exists limited information in these regions. The middle and bottom rows illustrate a failure mode consisting of very strong highlights that affect a large unsaturated region. Our method over-corrects the flares and leaves artifacts around the street lamps. This requires further exploration on extreme cases with large and strong highlights.

6. Visual Results on Simulated Dataset

In this section, we demonstrate additional visual results on simulated data. As shown in Figure 6, Figure 7, and Figure 8, the proposed DISCNet suppresses flare and haze effects around highlights, and removes most artifacts in nearby unsaturated regions.

7. Visual Results on Real Dataset

Post-processing. Since it is beyond the scope of this paper to build a full Image Signal Processor (ISP) to output final images from raw data, we only perform a simple post-processing pipeline on the input data to adjust the color intensities and approximate the color of camera outputs, which typically exhibit perceptually better color for viewing on a display. The post-processing includes 1) color correction by color correction matrix (CCM) from the camera, 2) RGB scaling which transforms camera RGB values into camera's output RGB values, and 3) contrast enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE). Note that we also adopt the post-processing pipeline to obtain similar color in the input images for visual comparisons.

Visual Comparisons. We provide more visual comparisons with representative methods on real data in Figure 9 and Figure 10. Our proposed network could remove diffraction image effects, while leaving least artifacts introduced by camera.

References

[1] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[2] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 267–276, 2002. 2

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[4] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 3

[5] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. *arXiv preprint arXiv:2003.04857*, 2020. 2, 3

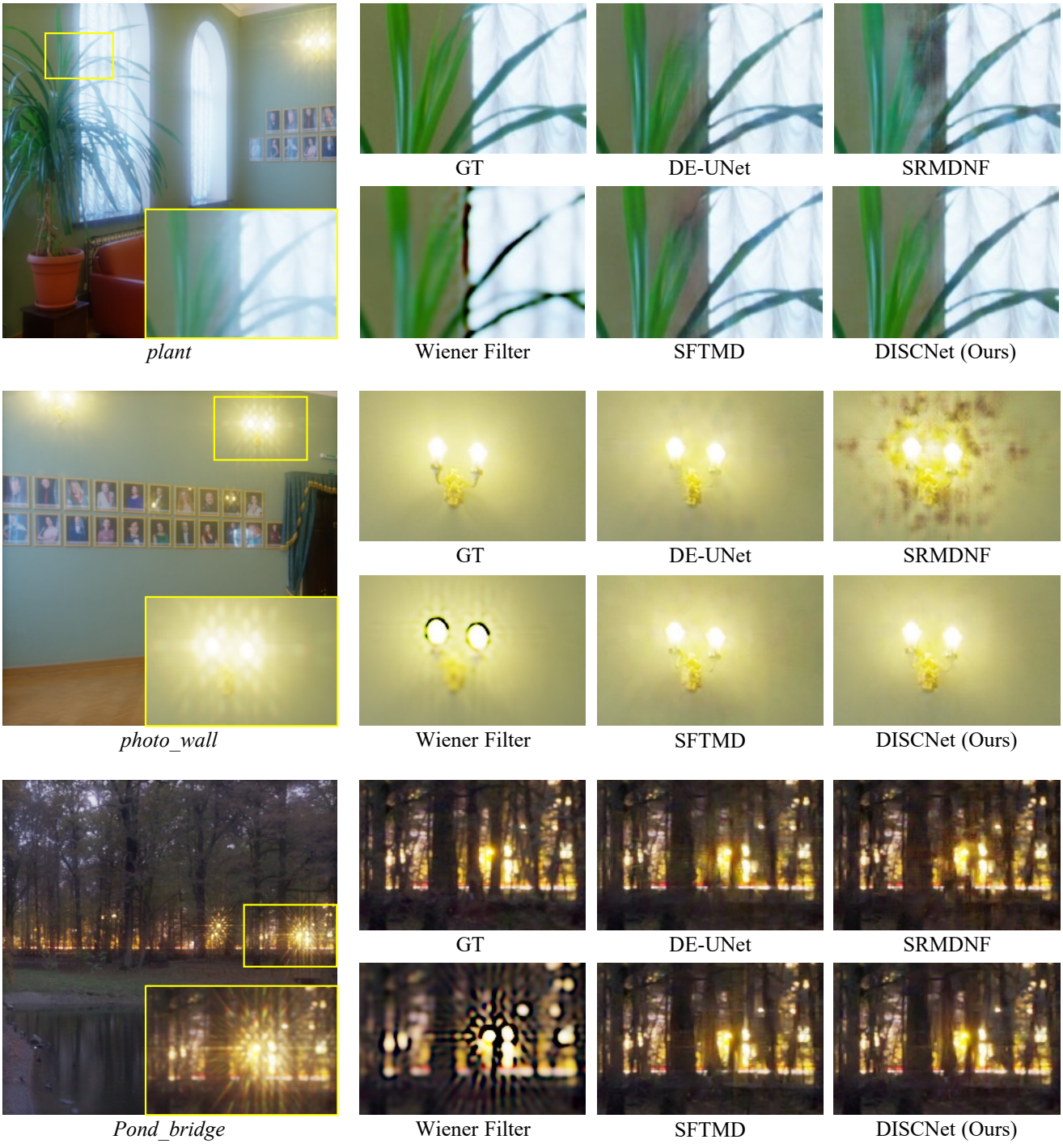


Figure 6. Visual comparison on simulated input images. (**Zoom in for better view.**)

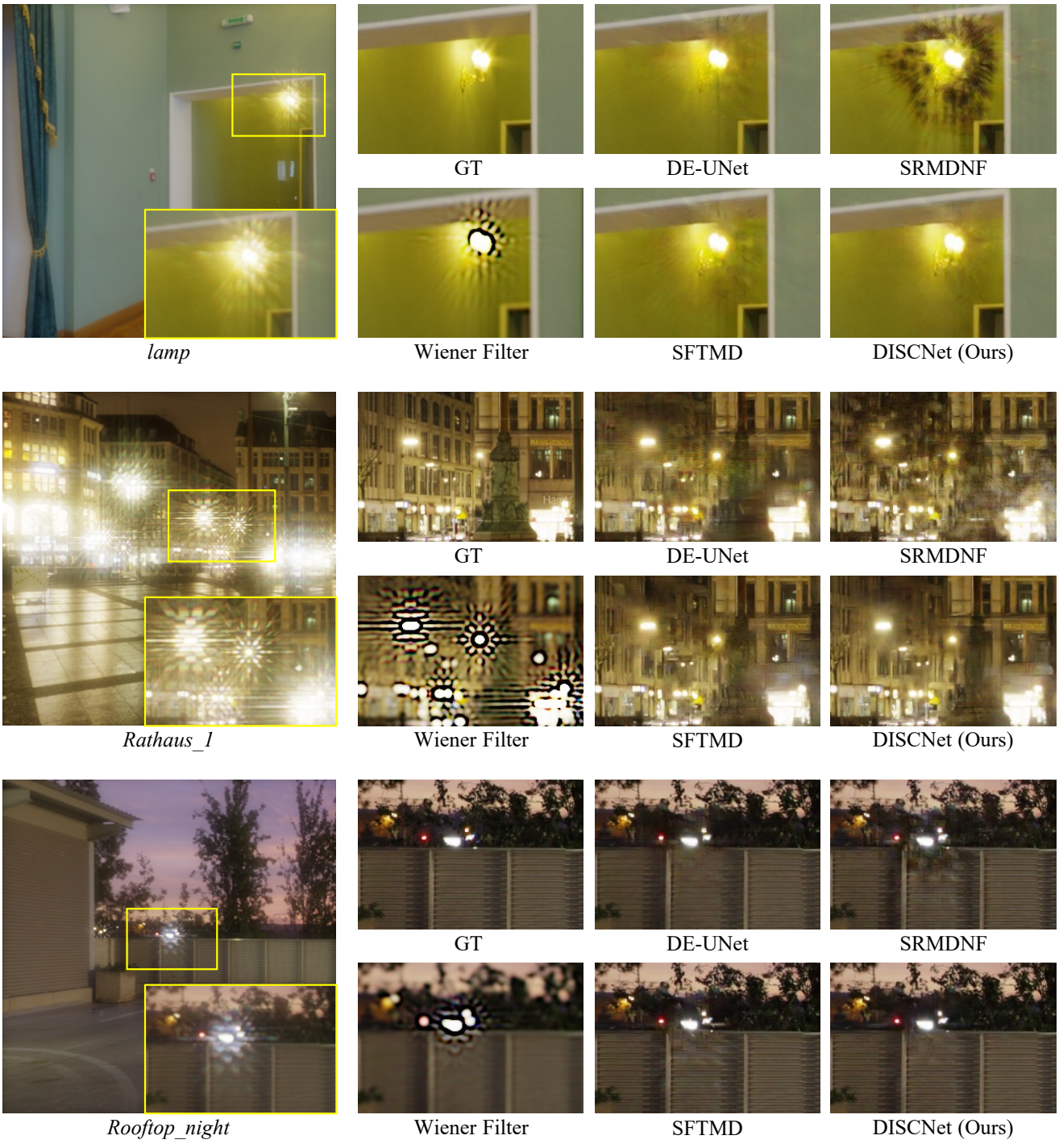


Figure 7. Visual comparison on simulated input images. (**Zoom in for better view.**)

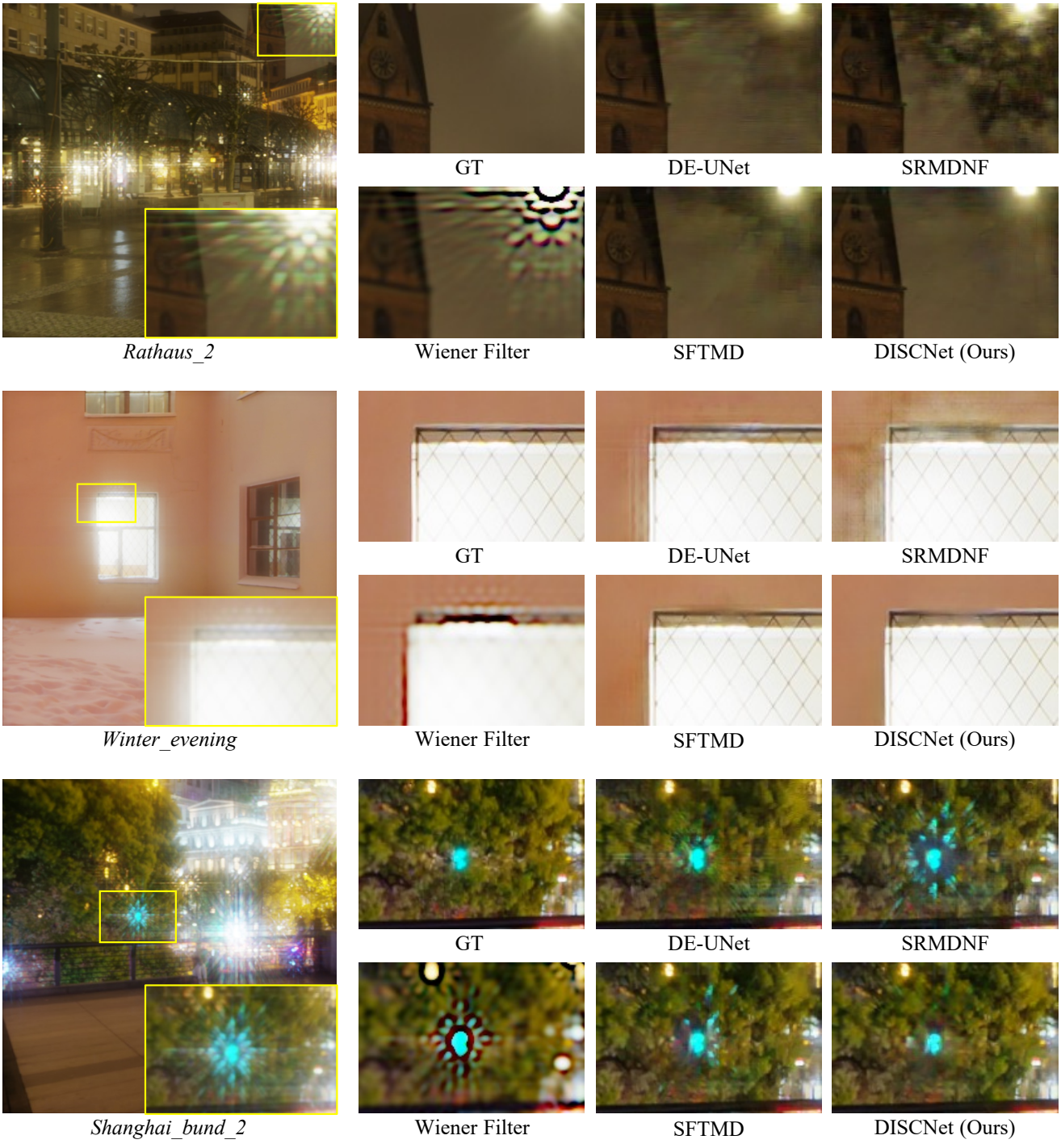


Figure 8. Visual comparison on simulated input images. (**Zoom in for better view.**)

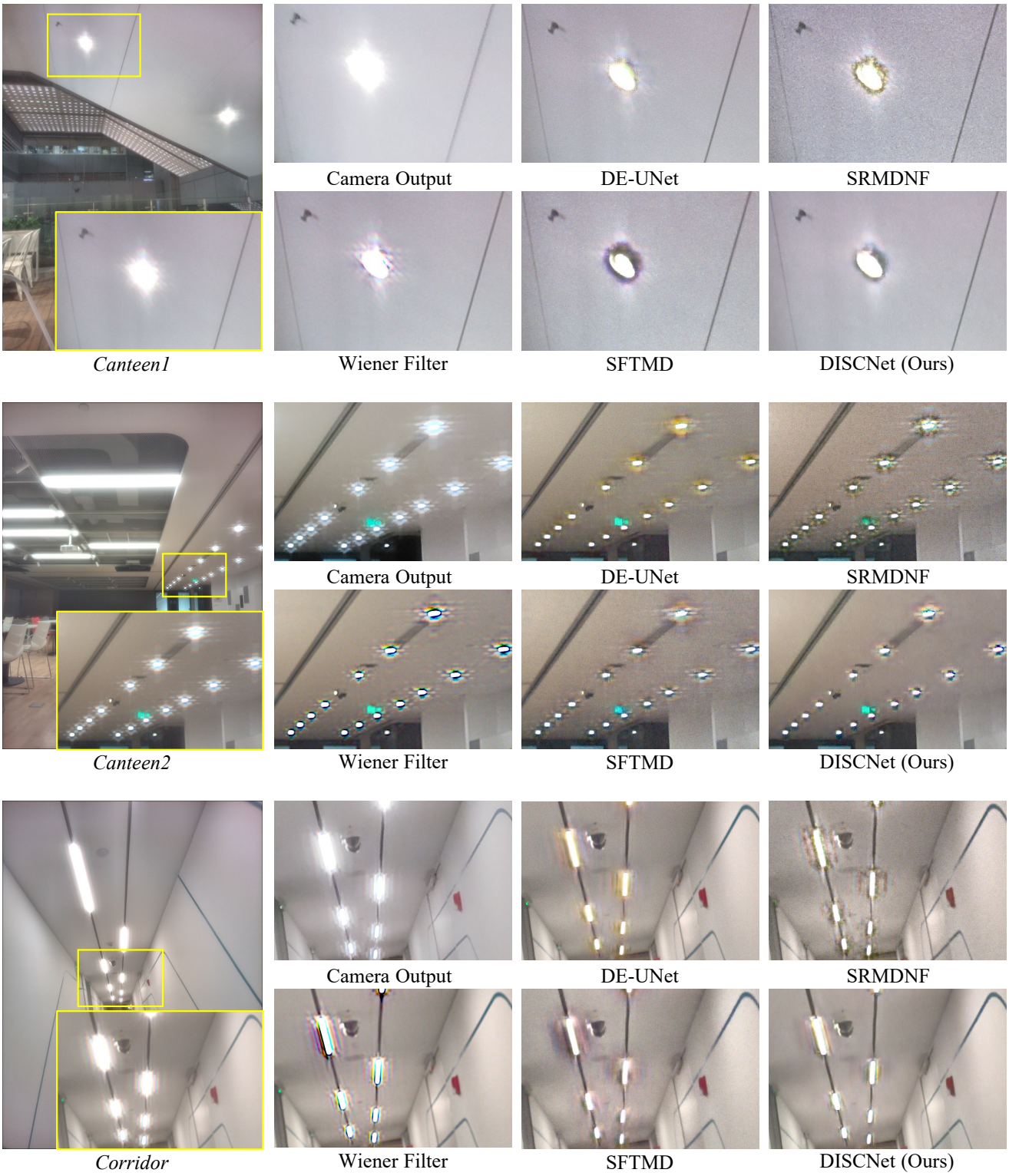


Figure 9. Visual comparison on real input images. (**Zoom in for better view.**)

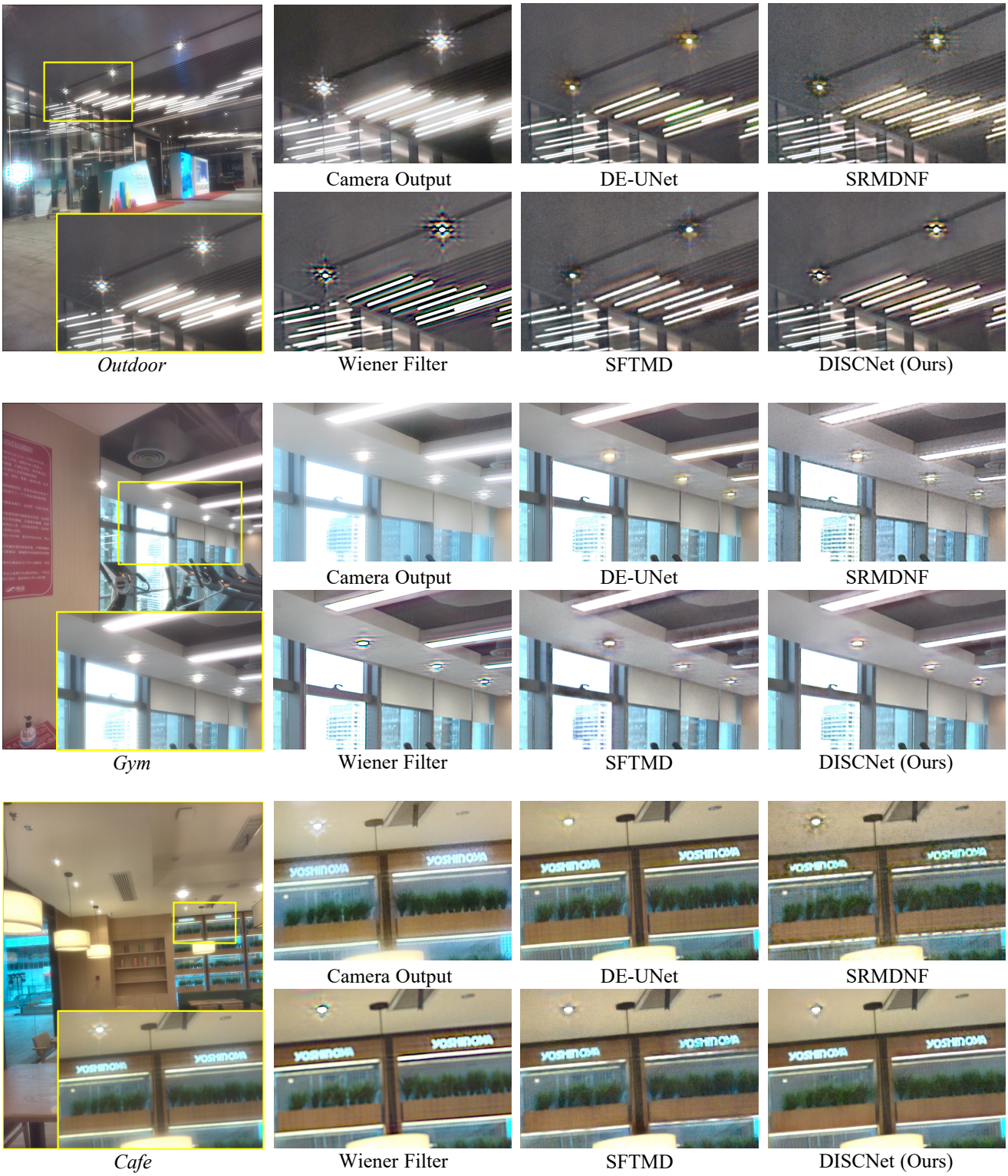


Figure 10. Visual comparison on real input images. (**Zoom in for better view.**)