[Supplementary Material]
FSI: Frequency and Spatial Interactive Learning for Image Restoration in Under-Display Cameras

Chengxu Liu[1,3], Xuan Wang[2], Shuai Li[2], Yuzhi Wang[2], Xueming Qian[1,3]
[1]Xi'an Jiaotong University [2]MEGVII Technology
[3]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd
liuchx97@gmail.com, {wangxuan02,lishuai,wangyuzhi}@megvii.com, qianxm@mail.xjtu.edu.cn

In this supplementary material, Sec. 1 provides the theoretical analysis of the distribution properties of diffraction in the frequency domain. Sec. 2 illustrates the detailed architecture of the color transform (CT) module. Sec. 3 describes more training details. Sec. 4 provides experiments to verify the effectiveness of the dual transfer unit (DTU). Sec. 5 shows spectral visualization comparisons. Sec. 6 analyzes the limitations of FSI. Finally, Sec. 7 shows more comparison results.

1. Theoretical Analysis

As described in the main paper, light emitted from a light source across a display with arranged organic light-emitting diodes (OLEDs) is diffracted, which will produce the periodic decreasing spectral biases captured by the sensor [5, 11]. In this section, we provide a detailed analysis of the theory involved.

Suppose there is a light emitted from a light source that crosses a display with arranged OLEDs. The effective aperture function $a(x,y)$ could be formulated as:

$$a(x,y) = g(x,y)p(x,y), \tag{1}$$

where $g(x,y)$ and $p(x,y)$ represent lens aperture and display openings, respectively. From basic Fourier optics, the point spread function (PSF), formulated as $k(x,y)$, can be written as the squared magnitude of the scaled Fourier transform of the effective aperture function, that's to say:

$$k(x,y) \propto \left|A(x,y)\right|^2, \tag{2}$$

where $A(x,y)$ is the Fourier transformation of the $a(x,y)$ in Eq. (1). In an under-display camera (UDC), since each display pixel is identical, the display openings are always periodic. If we denote $m(x,y)$ to be the opening pattern as pertaining to a single pixel, the overall display openings $p(x,y)$ can be constructed with copies of $m(x,y)$:

$$p(x,y) = m(x,y) * \sum^{r}\sum^{c}\delta(x-rT)\delta(y-cT), \tag{3}$$

where $\delta(x)$ is the Dirac delta function, r and c means row and col in lens aperture, the display pixels are repeating at a periodicity of T along both axes.

Since the fact that the power spectral density and auto-correlation are Fourier pairs, from Eq. (2), the Fourier transform of $k(x,y)$ can be expressed as an auto-correlation function of $a(x,y)$, formulated as:

$$K(x,y) = AC_a(x,y). \tag{4}$$

As shown in Eq. (1) and Eq. (3), the auto-correlation of $a(x,y)$ depends on the lens aperture $g(x,y)$ as well as the per-pixel display opening $m(x,y)$, when the pitch of the display T is significantly smaller than the lens aperture, there are multiple display pixels within the aperture. In this scenario, the auto-correlation AC_a at small displacements (x,y) becomes repeating copies of AC_m, the auto-correlation of $m(x,y)$, scaled by the number of copies of $m(x,y)$ within the lens aperture. The auto-correlations associated with T/P-OLED displays are shown in Fig. 4 in [11]. The periodic structures with peaks and nulls can be clearly observed due to a consequence of the periodicity of the display tiling. Here, we directly see the effect of the per-pixel pattern $m(x,y)$ and its periodic tiling in the invertibility of the PSF.

2. Architecture Details

As described in Sec. 3.5 of the main paper, degradation caused by multiple thin-film layers in the display usually produces color shifts. Therefore, inspired by the solutions in white balance [1, 4], we use a lightweight U-Net [6] to predict a set of coefficients to adjust the color temperature by matrix transformation.

In this section, we illustrate the detailed architecture of the lightweight U-Net as shown in Fig. 1. The entire network is based on the U-Net architecture, where the number of channels in each layer is marked. We follow the existing approach [3, 7] to extract features by stacking some Smoothed Dilated Residual Blocks on each scale. Each block contains atrous (or dilated) convolutions to expand the network's receptive field without loss in spatial resolution. Besides, inserting a convolutional layer before each dilation convolution enables computational and parameter efficiency via shared separable kernels. The parameters of the color transform modules are 1.03M.

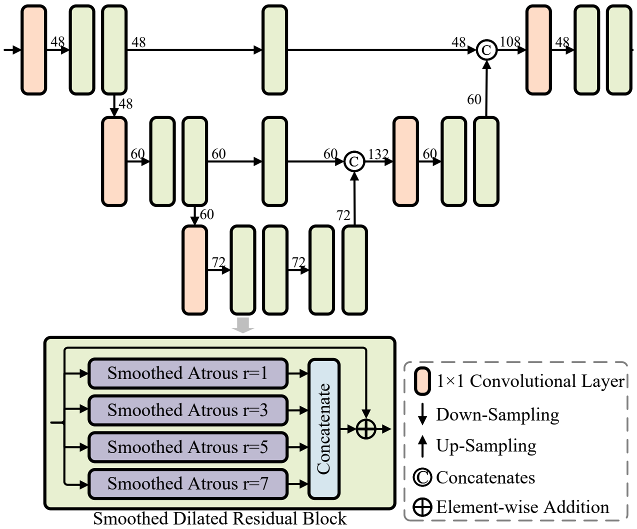


Figure 1. Network structure of color transform (CT) module.

3. More Training Details

As described in Sec. 3.3 of the main paper, the FSI consists of a series of stacked FSJ modules, and we grouped the two FSJ modules using a skip connection. To further improve the model performance, we add the loss function between the output and the ground truth after each group. This approach can effectively facilitate feature learning for each part.

4. Experimental Analysis of DTU

As described in Sec. 3.4 of the main paper, we present a Dual Transfer Unit (DTU) to enable selective interaction and joint learning of the frequency and spatial features within the modules. In particular, we first compress the features into a low-dimensional embedding space, and then learn them with convolution in the vertical and horizontal directions. To demonstrate the superiority of this approach, we compared other methods that have the ability to locate interactive features as shown in Tab. 1.

As shown in Tab. 1, our approach has higher performance for comparable parameters compared to the two typical attention mechanisms Spatial Attention [10] and DFC Attention [8]. Besides, since the self-attention mechanism in Transformer [9] relies on high-dimensional features to model the correlation between different regions, more parameters are required. Our method allows the network to facilitate the learning of complementary features with little cost and is much simpler than others.

Method	#P	PSNR	SSIM	LPIPS	DISTS
w/o	-	45.79	0.9935	0.0116	0.0183
Self-Attention [9]	10.9k	45.96	0.9936	0.0110	0.0152
Spatial Attention [10]	1.9k	45.96	0.9936	0.0110	0.0152
DFC Attention [8]	1.7k	46.00	0.9938	0.0110	0.0150
DTU(Ours)	1.7k	**46.05**	**0.9938**	**0.0109**	**0.0147**

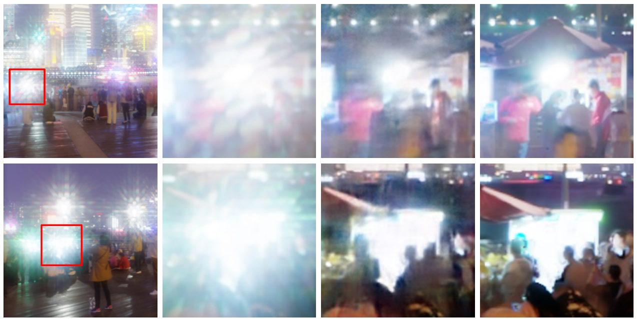
Table 1. Comparison between different methods of obtaining transfer gates in DTU on the SYNTH [2] dataset.

5. Spectral Visualization

To validate the capability of our method in reconstructing the frequency spectrums, we visualize the frequency spectrums of images and patches restored by different methods in Fig. 3. It can be observed that FSI has a great improvement in visual quality and the spectrums are also most similar to the ground truth. For example, in the first case, our FSI effectively eliminates the spectral bias generated by diffractions. The superior performance is owed to the learning capability of FSI in the frequency domain.

6. Limitation Analysis

Our work is superior in recovering regular textures (*e.g.*, the fourth row of Fig. 7 in the main paper) through frequency domain learning. However, for the recovery of the irregular texture loss in large areas, there are still some limitations. In this section, we visualize the failure cases of FSI in Fig. 2. Irregular textures over large areas are anomalous in frequency (*i.e.*,), so the reconstruction of the spectrum cannot effectively recover such content. It is worth noting that FSI still achieves greater gains than other methods through its powerful feature learning capability.



UDC Image FSI(Ours) Ground Truth

Figure 2. The failure cases when irregular texture loss occurs.

7. More Visualization Results

To further verify the effectiveness of our method, we show more comparison results among the proposed FSI and other advanced methods on three different benchmarks. The results on **P-OLED** [12], **T-OLED** [12], and **SYNTH** [2] are shown in Fig. 4, Fig. 5, and Fig. 6, respectively.

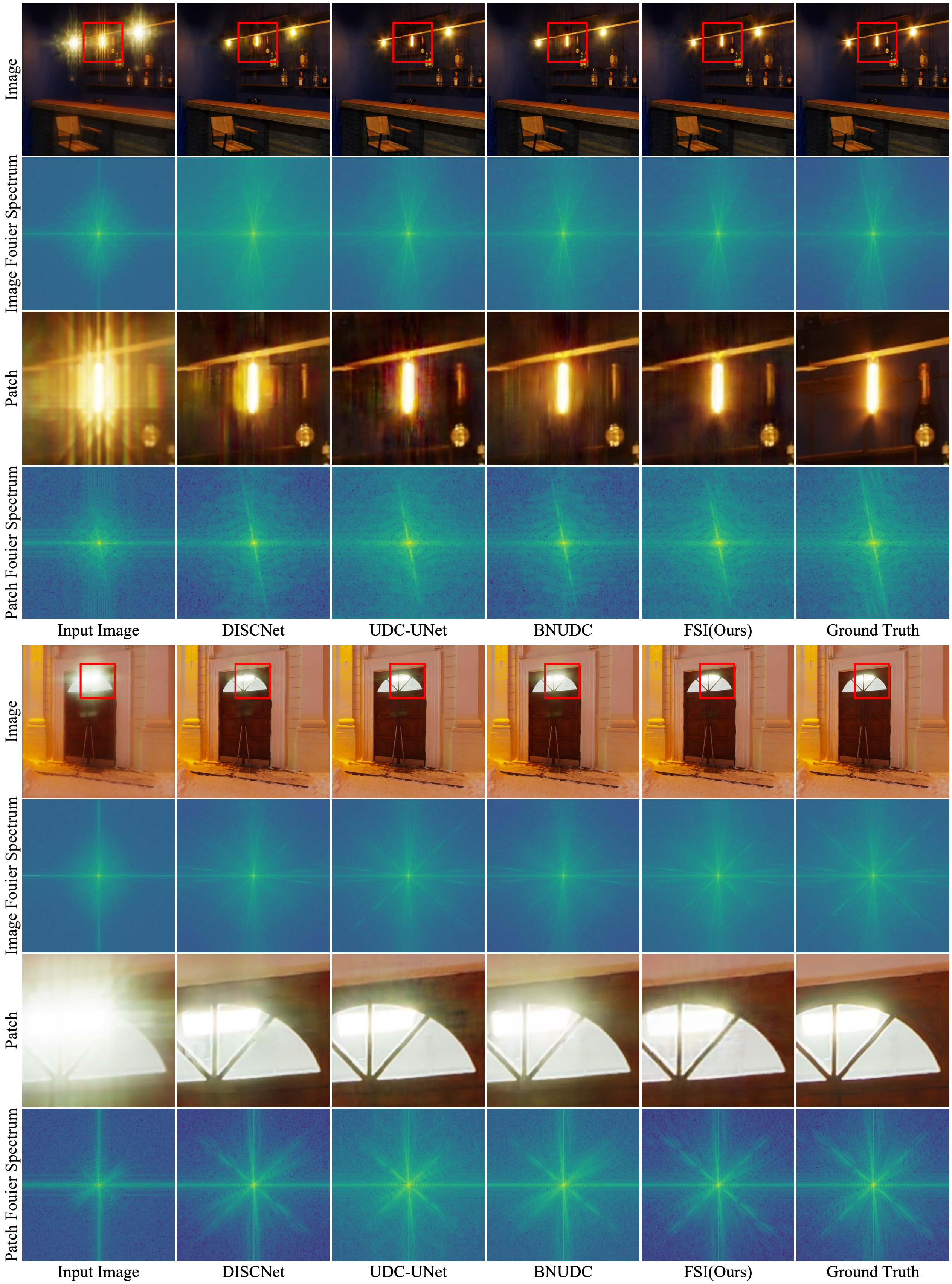


Figure 3. Visualization of the frequency spectrums for the entire image and patch regions. The method is shown at the bottom of each case.

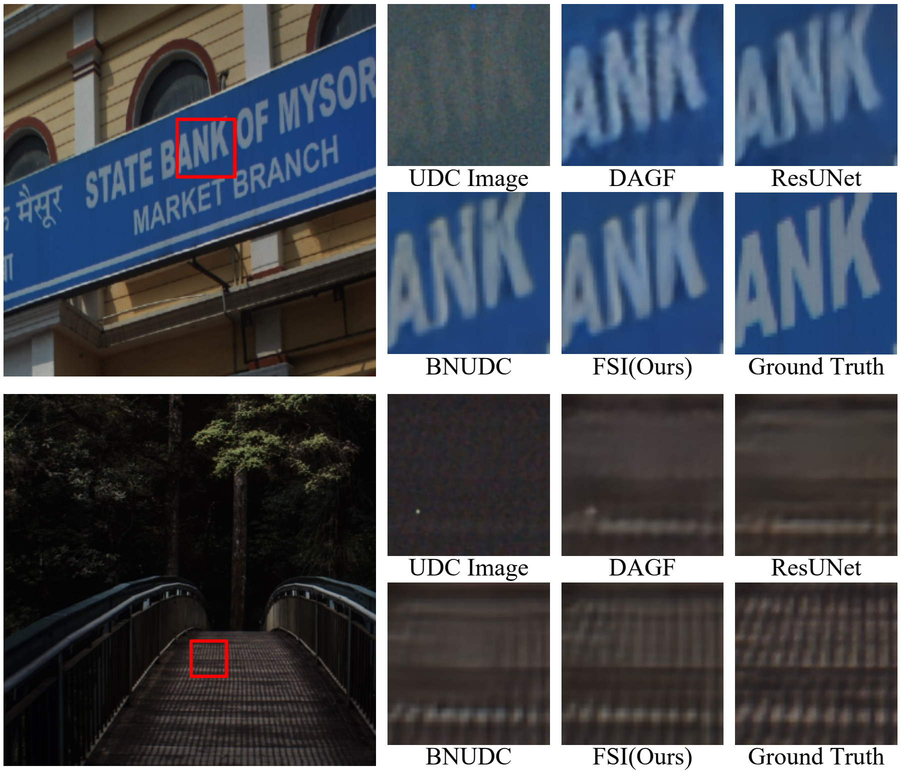


Figure 4. Visual results on P-OLED [12] dataset. The method is shown at the bottom of each case. Zoom in to see better visualization.

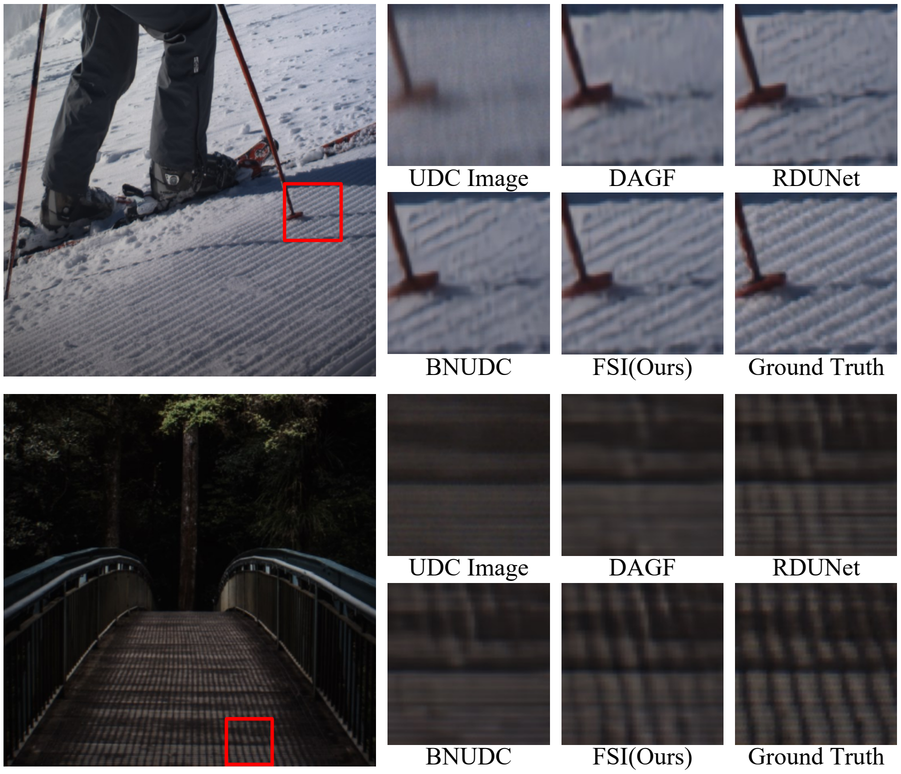


Figure 5. Visual results on T-OLED [12] dataset. The method is shown at the bottom of each case. Zoom in to see better visualization.

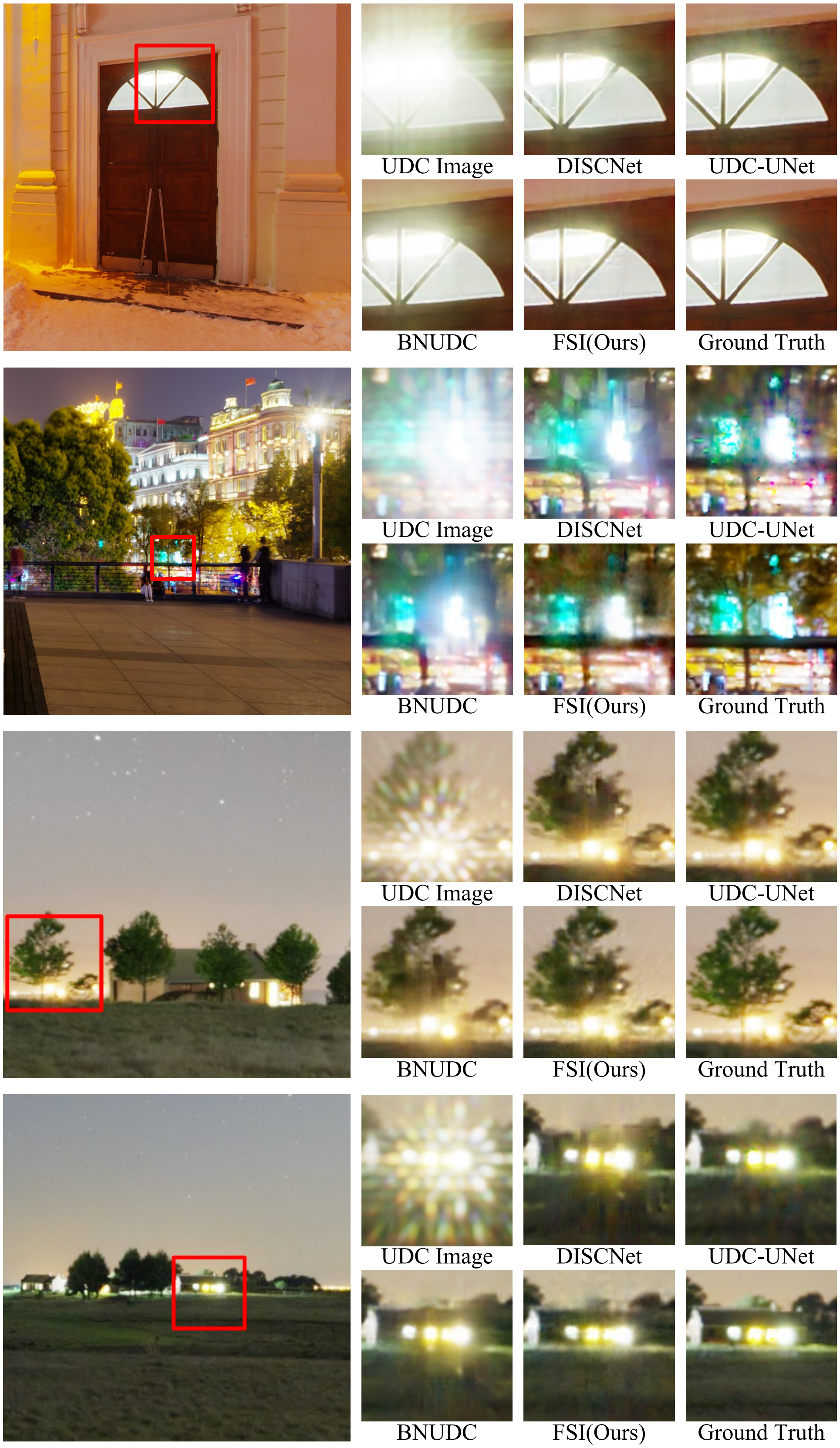


Figure 6. Visual results on SYNTH [2] dataset. The method is shown at the bottom of each case. Zoom in to see better visualization.

References

[1] Mahmoud Afifi and Michael S Brown. Deep white-balance editing. In *CVPR*, pages 1397–1406, 2020. 1

[2] Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, and Jinwei Gu. Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In *CVPR*, pages 662–671, 2021. 2, 3, 6

[3] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. Bnudc: A two-branched deep neural network for restoring images from under-display cameras. In *CVPR*, pages 1950–1959, 2022. 2

[4] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4D LUT: Learnable context-aware 4d lookup table for image enhancement. *arXiv preprint arXiv:2209.01749*, 2022. 1

[5] Zong Qin, Yu-Hsiang Tsai, Yen-Wei Yeh, Yi-Pai Huang, and Han-Ping David Shieh. See-through image blurring of transparent organic light-emitting diodes display: calculation method based on diffraction and analysis of pixel structures. *Journal of Display Technology*, 12(11):1242–1249, 2016. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1

[7] Varun Sundar, Sumanth Hegde, Divya Kothandaraman, and Kaushik Mitra. Deep atrous guided filter for image restoration in under display cameras. In *ECCVW*, pages 379–397. Springer, 2020. 2

[8] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Chao Xu, and Yunhe Wang. GhostNetV2: Enhance cheap operation with long-range attention. *NeurIPS*, 2022. 2

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 2

[11] Anqi Yang and Aswin C Sankaranarayanan. Designing display pixel layouts for under-panel cameras. *IEEE TPAMI*, 43(7):2245–2256, 2021. 1

[12] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *CVPR*, pages 9179–9188, 2021. 3, 5