Generating Aligned Pseudo-Supervision from Non-Aligned Data for Image Restoration in Under-Display Camera - Supplementary Material -

Ruicheng Feng[1] Chongyi Li[1] Huaijin Chen[2] Shuai Li[2] Jinwei Gu[3,4] Chen Change Loy[1]
[1]S-Lab, Nanyang Technological University [2]SenseBrain Technology
[3]The Chinese University of Hong Kong [4]Shanghai AI Laboratory
{ruicheng002, chongyi.li, ccloy}@ntu.edu.sg
{huaijin.chen, shuailizju}@gmail.com jwgu@cuhk.edu.hk

A. Implementation Details

A.1. Camera Setup

To build a real-world dataset, we mount the two smartphones on the tripod as shown in Figure A1 and take each set of images using Bluetooth remote controller to control the shutter speed. The camera modules are physically placed as close as possible to decrease the baseline in a stereo setting. The aperture and focal length are fixed and unadjustable for both cameras. The resolution of ZTE Axon 20 is 3264×2448, and that of iPhone 13 Pro is 4032×3024. For the *degraded image*, we set the under-display camera configurations by its built-in automatic exposure system and take three shots bracketed at $[1, 1/4, 1/16]$, which are then composed into one HDR image. For the *reference image*, we set a low ISO value ranging from 100 to 200 to avoid heavy noise, and adjust the shutter speed to capture sharp and clean images of proper exposure. We avoid capturing objects that are too close to filter out image pairs with large parallax or occlusion. For each scene of the pair, we register the image captured by iPhone as a reference image to the UDC image using a homography transform calculated by RANSAC [16]. Then we crop out the invalid areas due to the homography transformation and parallax between two cameras, and downsample the pairs to 3200×2400. The dataset is exemplified by triplet sets in Figure A2. The reference images after alignment still show mild displacement compared to the corresponding UDC images.

A.2. Occlusion Mask

Occluded pixels, by definition, are invisible in the reference image, which should be discounted for spatial supervision (*e.g.*, $\mathcal{L}_1$ and $\mathcal{L}_{VGG}$), since inaccurate deformations over these regions could deteriorate image restoration network training. We implement the detection based on forward-backward consistency assumption [13], that is, for non-occluded pixels, traversing the forward flow and then backward should arrive at the same pixel.

Figure A1. A close-up picture of our custom-built camera setup consisting of two smartphones mounted on a tripod.

In particular, the forward optical flow from I_R to $\hat{I_D}$ is given by $\Psi^f = \psi_{flow}(\hat{I_D}, I_R)$, and the backward flow can similarly be estimated by $\Psi^b = \psi_{flow}(I_R, \hat{I_D})$. Suppose forward flow vector is $\mathbf{w} = \Psi^f(\mathbf{p})$ at point $\mathbf{p} = (x, y)$, the forward-backward flow vector is denoted as $\hat{\mathbf{w}} = \Psi^b(\mathbf{p} + \Psi^f(\mathbf{p}))$. The non-occlusion mask M, with granted tolerance for small estimation errors, can then be formulated by

$$M(\mathbf{p}) = \begin{cases} 1, & \text{if } \|\mathbf{w} + \hat{\mathbf{w}}\|_2 < \alpha(\|\mathbf{w}\|_2 + \|\hat{\mathbf{w}}\|_2) + \beta \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where tolerance parameters α and β are set to 0.1 and 1 in this paper. Figure A3 illustrates several examples of masked invalid regions (highlighted in red).

A.3. Conditional PatchGAN

To further improve the visual quality, we also add adversarial loss based on conditional PatchGAN [5]:

$$\mathcal{L}_{GAN} = -\mathbb{E}[\log \mathcal{D}(I_D, I_O)], \tag{2}$$

Figure A2. **The image examples in our dataset.** Images (a) and (b) are captured by UDC and a normal camera, and they exhibit obvious misalignment. We roughly align (b) to (a) with homography transformation to obtain (c). Even after alignment, there still exists mild misalignment between (a) and (c), which will be mitigated by our AlignFormer.

where $\mathcal{D}$ is the discriminator conditioned on both input and output of PPM-UNet, and it is optimized by

$$\mathcal{L}_{\mathcal{D}} = -\mathbb{E}[\log\mathcal{D}(I_D, I_P)] - \mathbb{E}[\log(1 - \mathcal{D}(I_D, I_O))]. \quad (3)$$

PatchGAN uses a discriminator that distinguishes image patches of size 70×70, which proves to produce sharper details than the vanilla "global" discriminator. Inspired by [8, 9], we adopt conditional PatchGAN as a discriminator to capture high frequencies in local features. Particularly, we use the discriminator architecture that only penalizes structure at the scale of patches (PatchGAN). This discriminator tries to classify if each $N \times N$ patch is real or fake. We run the discriminator across the image, and then average all responses to obtain the patch-based output. N is set to 16 in our experiments. We also empirically found conditional GAN, where the discriminator is conditioned on UDC images, facilitates more realistic results.

A.4. Network Structure

The Domain Alignment Module (DAM) contains two sub-nets: a guidance net and a matching net. The detailed architecture of DAM is listed in Table A1. In the matching net, StyleConv is the conv layer modulated by style conditions as proposed in StyleGANv2 [7]. The guidance net is designed to generate a conditional vector that extracts holistic domain information from the reference image. After that, the matching net transfers the domain information, *e.g.*, color, illuminance, and contrast, to the degraded UDC image and produces a coarse restored image that is similar to the reference image.

The structure of the image restoration network is shown in Table A3. We adopt a modified U-Net as in [20] and add a Pyramid Pooling Module (PPM) [19] into the network to capture global information (See Figure A4). We adopt the original design of PPM containing 4 mean pooling branches

Figure A3. **Invalid mask visualization.** The invalid (occlusion) mask (highlighted in red) is estimated by forward-backward assumption. Note that it also detects pixels at borders, and moving objects in non-still images.

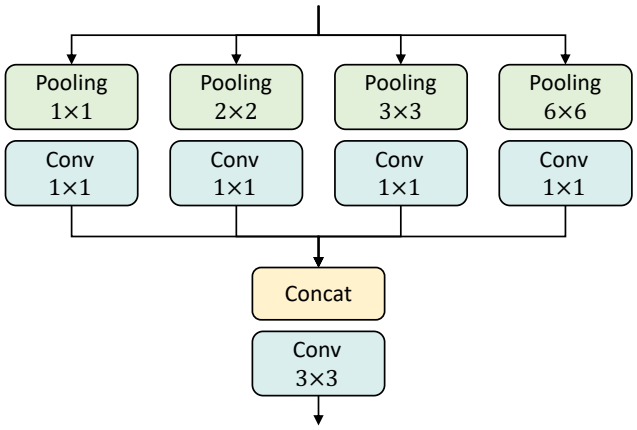


Figure A4. **Illustration of Pyramid Pooling Module (PPM).**

with bin sizes of 1, 2, 3, 6. As demonstrated in the main paper, the PPM layers propagate global prior into the network and stabilize training and suppress artifacts in UDC image restoration.

B. Objective Evaluation

In addition to quantitatively measuring the characteristics of the captured natural scenes (*i.e.*, on the test dataset), we also conduct objective quality evaluation via engineering-based quality metrics. The Modulation Transfer Function (MTF) and the related Spatial Frequency Response (SFR) are commonly used to characterize an imaging system's reproduction of modulation, as a function of spatial frequency [11, 15, 17]. MTF can inform the system's resolution and sharpness, which mainly contribute to the overall image quality. MTF can be derived from the slanted-

Table A1. **Detailed structure of DAM.** k and s indicate the kernel size and stride of the convolutional layer. $\uparrow$ and $\downarrow$ represent $2\times$ upsampling and $2\times$ downsampling, respectively.

Guidance Net		
Layer	Configuration	Output size
Conv, LeakyReLU	$k=3, s=1$	$256\times256\times64$
Conv, LeakyReLU	$k=3, s=2$	$128\times128\times64$
Conv, LeakyReLU	$k=3, s=1$	$128\times128\times64$
Conv, LeakyReLU	$k=3, s=1$	$64\times64\times64$
Conv	$k=3, s=1$	$64\times64\times64$
Global Average Pooling	-	$1\times1\times64$
Matching Net		
Layer	Configuration	Output size
StyleConv, LeakyReLU	$k=3, s=1$	$256\times256\times64$
StyleConv, LeakyReLU, $\downarrow$	$k=3, s=1$	$128\times128\times64$
StyleConv, LeakyReLU	$k=3, s=1$	$128\times128\times64$
StyleConv, LeakyReLU, $\downarrow$	$k=3, s=1$	$64\times64\times64$
StyleConv, LeakyReLU	$k=3, s=1$	$64\times64\times64$
StyleConv, LeakyReLU	$k=3, s=1$	$64\times64\times64$
StyleConv, LeakyReLU	$k=3, s=1$	$64\times64\times64$
StyleConv, LeakyReLU, $\uparrow$	$k=3, s=1$	$128\times128\times64$
StyleConv, LeakyReLU	$k=3, s=1$	$128\times128\times64$
StyleConv, LeakyReLU, $\uparrow$	$k=3, s=1$	$256\times256\times64$
StyleConv, LeakyReLU	$k=3, s=1$	$256\times256\times64$
StyleConv, LeakyReLU	$k=3, s=1$	$256\times256\times3$

Table A2. **The MTF curve results.** Reported are the weighted mean summary of several detected slant edges.

Metric	UDC	Ref	Restored
MTF50 (LW/PH) $\uparrow$	661	1516	1039
MTF20 (LW/PH) $\uparrow$	1307	2023	1769

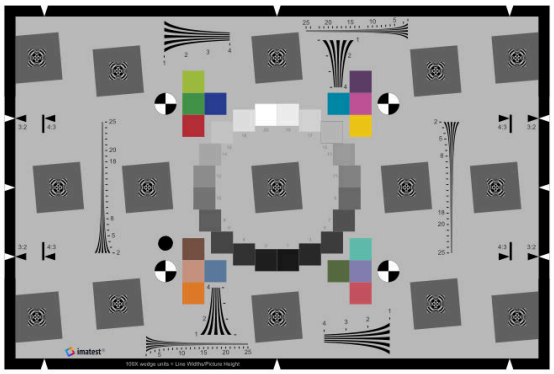


Figure A5. Lab-based image of ISO 12233 eSFR test chart.

Table A3. **Detailed structure of PPM-UNet.**

Module	Kernel size	# of channels	Dilation	Stride	Activation	Output size
Conv1	3×3	32	1	1	LeakyReLU (0.2)	$256 \times 256 \times 32$
Conv2	3×3	64	1	2	LeakyReLU (0.2)	$128 \times 128 \times 64$
Conv3	3×3	64	1	1	LeakyReLU (0.2)	$128 \times 128 \times 64$
PPM1	-	-	-	-	-	$128 \times 128 \times 64$
Conv4	3×3	128	1	2	LeakyReLU (0.2)	$64 \times 64 \times 128$
Conv5	3×3	128	1	1	LeakyReLU (0.2)	$64 \times 64 \times 128$
PPM2	-	-	-	-	-	$64 \times 64 \times 128$
Conv6	3×3	128	1	2	LeakyReLU (0.2)	$32 \times 32 \times 128$
Conv7	3×3	128	1	1	LeakyReLU (0.2)	$32 \times 32 \times 128$
PPM3	-	128	-	-	-	$32 \times 32 \times 128$
Conv8	3×3	128	1	1	LeakyReLU (0.2)	$32 \times 32 \times 128$
Conv9	3×3	128	1	1	LeakyReLU (0.2)	$32 \times 32 \times 128$
Add (w/ PPM3)	-	128	-	-	-	$32 \times 32 \times 128$
Upsample $\uparrow$	-	128	2	-	-	$64 \times 64 \times 128$
Conv10	3×3	128	1	1	LeakyReLU (0.2)	$64 \times 64 \times 128$
Conv11	3×3	128	1	1	LeakyReLU (0.2)	$64 \times 64 \times 128$
Add (w/ PPM2)	-	128	-	-	-	$64 \times 64 \times 128$
Upsample $\uparrow$	-	128	2	-	-	$128 \times 128 \times 128$
Conv12	3×3	64	1	1	LeakyReLU (0.2)	$128 \times 128 \times 64$
Conv13	3×3	64	1	1	LeakyReLU (0.2)	$128 \times 128 \times 64$
Add (w/ PPM1)	-	64	-	-	-	$128 \times 128 \times 64$
Upsample $\uparrow$	-	64	2	-	-	$256 \times 256 \times 64$
Conv14	3×3	32	1	1	LeakyReLU (0.2)	$256 \times 256 \times 32$
Conv15	3×3	32	1	1	LeakyReLU (0.2)	$256 \times 256 \times 32$
Conv16	3×3	3	1	1	-	$256 \times 256 \times 3$

edge technique [2] with carefully designed test charts under strict laboratory conditions. To calculate the MTF curve, we use the Enhanced version of ISO 12233 imatest eSFR test chart [4] (See Figure A5). The enhanced eSFR ISO test chart adds 6 squares on sides, 16 color patches, and several wedge patterns.

A key data point from the MTF curve is MTF50, where MTF is 50% of its low (0) frequency value. Similarly, MTF20 is the spatial frequency where MTF is 20% of the zero frequency. As recommended by Imatest [1], we use these two metrics for analysis throughout this work. We use the data dumps with least post-processing, such that the evaluation can operate in a more linear region, and hence results are less affected by overexposure, underexposure, and excessive sharpening. Table A2 summarizes the results. As can be observed, the MTF values increase when the UDC images are restored by our PPM-UNet, which demonstrates that the image restoration network also increases the contrast and sharpness of images.

In addition, we show the complete MTF curves, and the edge profile the MTF is derived from, of the original UDC, reference, and restored UDC in Figure A6. One can observe the contrast improvement in the low- to mid-frequency band of the restored image compared to the original UDC image, as the modulation transfer is noticeably higher in the 0-0.4 cycles/pixel frequency region. The overall MTF shape of the restored image is also more similar to the reference image compared to the original UDC, suggesting an overall more natural contrast and sharpness after restoration.

C. Analysis of Displacement Metrics

In absence of ground-truth correspondence, it is non-trivial to quantify how well the pseudo GT (output of AlignFormer) is aligned to the UDC image. Thus, we indirectly measure the displacement error with LoFTR [12] that serves as a keypoint matcher. Given a set of matched keypoints from two images, PCK measures the Percentage of Correct Keypoints transferred to another image, which lie within a

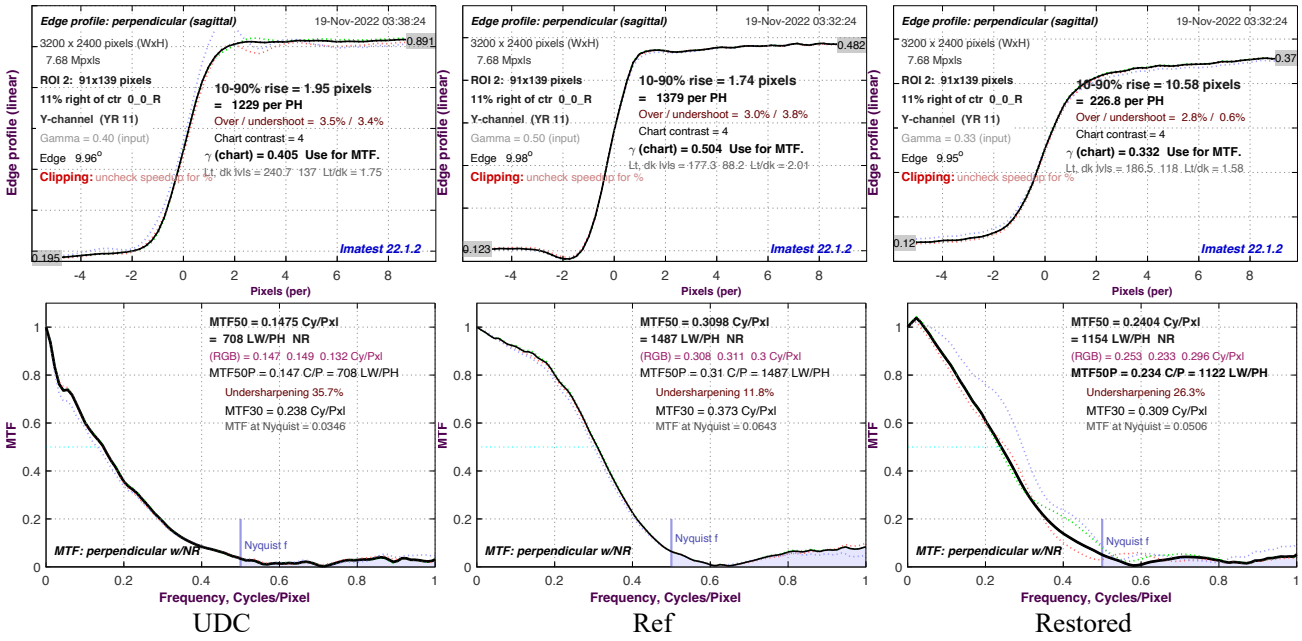


Figure A6. MTF curves and edge profiles, obtained with Imatest on ISO12233 test chart.

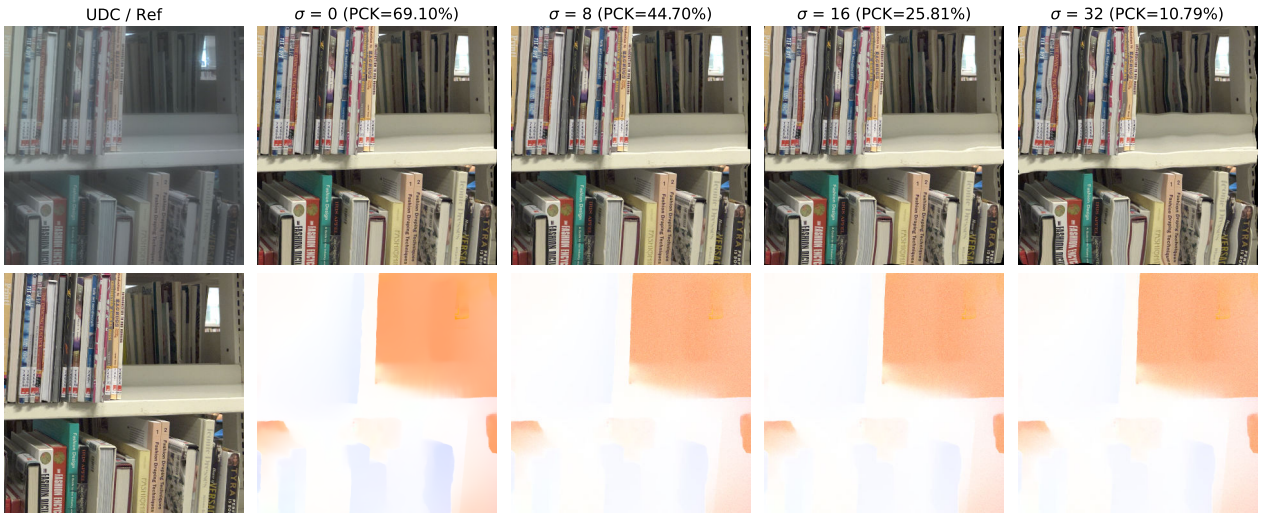


Figure A7. **Illustration of PCK and flows with perturbation.** Top row are the image and wrapped images, while bottom row show the flows with perturbation. The wrapped images are severely distorted when injecting large perturbation on flow. Reported results are evaluated on PCK with $\alpha = 0.01$ ($\sim$ 1 pixel error tolerance).

certain error threshold.

In particular, suppose $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are the same matched keypoint located at $\mathbf{p}_A$ in one image and $\mathbf{p}_B$ in another image, $d = \|\mathbf{p}_A - \mathbf{p}_B\|_2$ measures the displacement of coordinates. Ideally, the offset should be all zeros when two images are perfectly aligned. As global keypoint search may lead to long-range matching, causing outliers with large displacement, we do not average out all displacement errors d over detected keypoints. Instead, we calculate the percentage of correct matched pairs, denoted as *PCK* in the main paper. The keypoint pair is deemed correctly aligned when d is smaller than the preset threshold. Following common practice in image matching [6, 10], we set α error threshold, given by $d < \alpha \times \max(H, W)$, where H and W are the height and width of the image.

To justify the metrics, we conduct controlled experiments for further analysis. Specifically, we estimate optical flow from the reference image to the UDC image us-

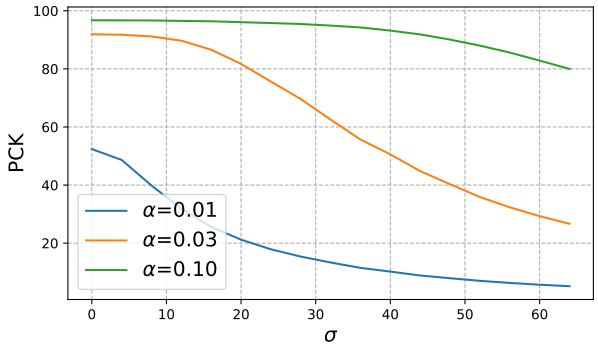


Figure A8. **PCK on different wrapped images where flows are added with various perturbations.** The perturbation is controlled by Gaussian noise with various std σ ranging from 0 (Original flow) to 64. We report the PCK results averaged over 330 images of shape 1024×1024. The PCK curves exhibit a consistent tendency with perturbation, indicating the displacement metric PCK can accurately reflect misalignment in our task.

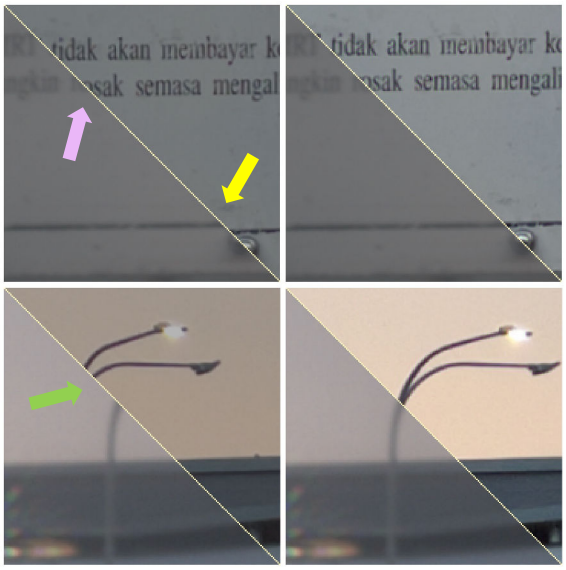
ing RAFT [14], inject perturbation into the flow, and then warp the reference image with the deformed flow. Following [18], the deformation is implemented by sampling amounts of independent Gaussian noise with $\mu = 0$ and controllable standard deviation (std) σ. Then we compute PCK between the warped reference image and UDC image. The reported PCK results are averaged over 330 images of size 1024×1024. Generally, a larger perturbation to the flow would induce greater displacement on the warped image. Figure A8 reflects the behaviors of PCK at various α thresholds under different perturbations. All curves consistently drop when the perturbations (controlled by σ) become greater. Figure A7 visualizes an example of warping flow with perturbation. We also observe a similar tendency of PCK and perturbation. This implies the PCK metric is suitable in our cases for quantifying misalignment.

D. Additional Ablation Studies

Effectiveness of Alignment Method. Figure 7 in the main body presents the results aligned by the image registration algorithm proposed by Cai *et al*. [3]. Since originally designed for image pairs taken at different focal lengths, it is hard to register stereo pairs where optical axis are not coincident. This problem is further compounded by the unique degradation of UDC images. As shown in Figure A9, Cai *et al*. cannot achieve accurate registration, while our AlignFormer perfectly aligns the reference images.

E. Additional Visual Results

We provide more visual comparisons on real data in Figure A10 and Figure A11. In addition, Figure A12 and A13 present more visual comparisons with representative works.



UDC – Cai *et al.* UDC – AlignFormer

Figure A9. **Stitched pairs of local patches from aligned images.** Bottom left are UDC images, and top right are corresponding results generated by Cai *et al*. [3] or AlignFormer. Results demonstrate Cai *et al*. [3] cannot achieve accurate registration, while our AlignFormer perfectly aligns the reference images.

Our method outperforms previous approaches in both removing artifacts and suppressing flare. Other methods fail to remove complicated artifacts or introduce over-correct artifacts), or produce blurry results. The visual results suggest that our proposed data generation framework could facilitate diffraction removal and restore texture details well.

F. Limitations

Although the PPM-UNet, as a baseline model, already achieves promising performance for restoring real UDC images, more domain-specific designs such as aiming at the limited dynamic range of UDC images and remedying the loss of details around the over-saturation regions are required for better restoration. While achieving rather satisfactory results on small areas of light sources, our work still struggles when highlight regions are large and intensities are extremely strong, leading to blurry results. This requires further exploration on extreme cases with large and strong highlights.

G. Broader Impacts

Our work can be used to generate other types of aligned pseudo-supervision from non-aligned data *e.g*. low-light/normal-light data and low-resolution/high-resolution data. Our data will enable neural networks for restoring UDC images. Our method will provide a new solution to

such issues for both academia and industry. While collecting our dataset, we try to avoid people to ensure privacy. Therefore, our dataset does not involve ethical issues. Moreover, as a typical image restoration task, our work will not bring negative impacts to the society.

References

[1] Iso 12233 — resolution and spatial frequency responses. https://www.imatest.com/solutions/iso-12233/. Accessed: 2022-11-18. 4

[2] Peter D Burns et al. Slanted-edge mtf for digital camera and scanner analysis. In *Is and Ts Pics Conference*, pages 135–138. Society for Imaging Science & Technology, 2000. 4

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 6

[4] International Organization for Standardization. *ISO 12233:2017 Photography - Electronic still picture imaging - Resolution and spatial frequency responses*. 2017. 4

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[6] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence Transformer for Matching Across Images. In *ICCV*, 2021. 5

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[8] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 2

[9] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *CVPR*, 2019. 2

[10] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *CVPR*, 2022. 5

[11] Joshua D Rego, Huaijin Chen, Shuai Li, Jinwei Gu, and Suren Jayasuriya. Deep camera obscura: an image restoration pipeline for pinhole photography. *Optics Express*, 30(15):27214–27235, 2022. 3

[12] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021. 4

[13] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010. 1

[14] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 6

[15] Oliver van Zwanenberg, Sophie Triantaphillidou, Robin Jenkin, and Alexandra Psarrou. Edge detection techniques for quantifying spatial imaging system performance and image quality. In *CVPRW*, 2019. 3

[16] Andrea Vedaldi and Brian Fulkerson. VLFeat: An open and portable library of computer vision algorithms. In *ACMMM*, 2010. 1

[17] D Vint, G Di Caterina, JJ Soraghan, RA Lamb, and D Humphreys. Evaluation of performance of vdsr super resolution on real and synthetic images. In *SSPD*, 2019. 3

[18] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *ECCV*. Springer, 2020. 6

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[20] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, 2022. 2

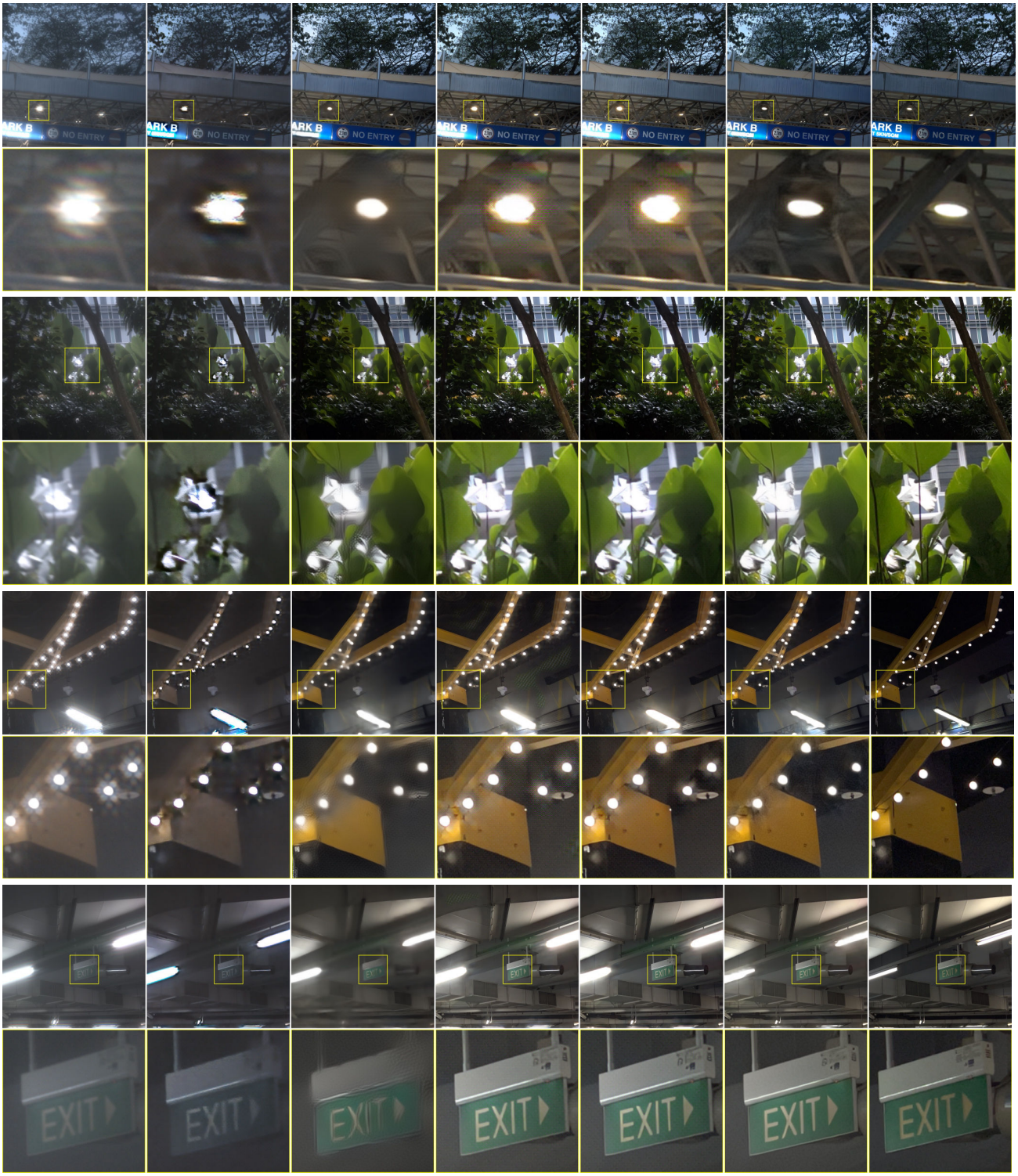
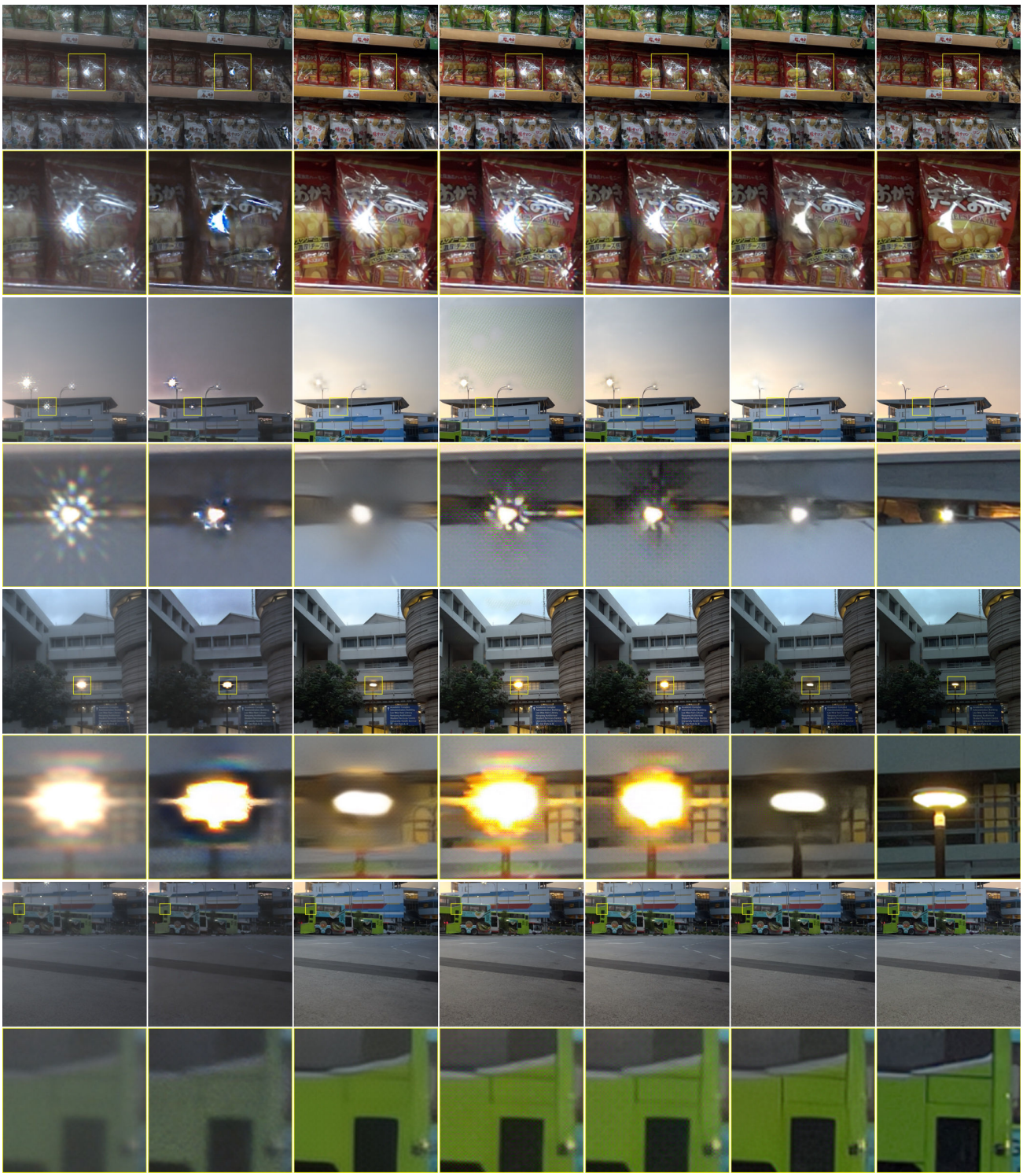


Figure A10. **Visual comparison between different datasets on the baseline network.**



UDC | *Synthetic* + $\mathcal{L}_{rec}$ | *Real* + $\mathcal{L}_{rec}$ | *Real* + $\mathcal{L}_{CX}$ | *Real* + $\mathcal{L}_{CoBi}$ | **Ours** | Ref

Figure A11. **Visual comparison between different datasets on the baseline network.**

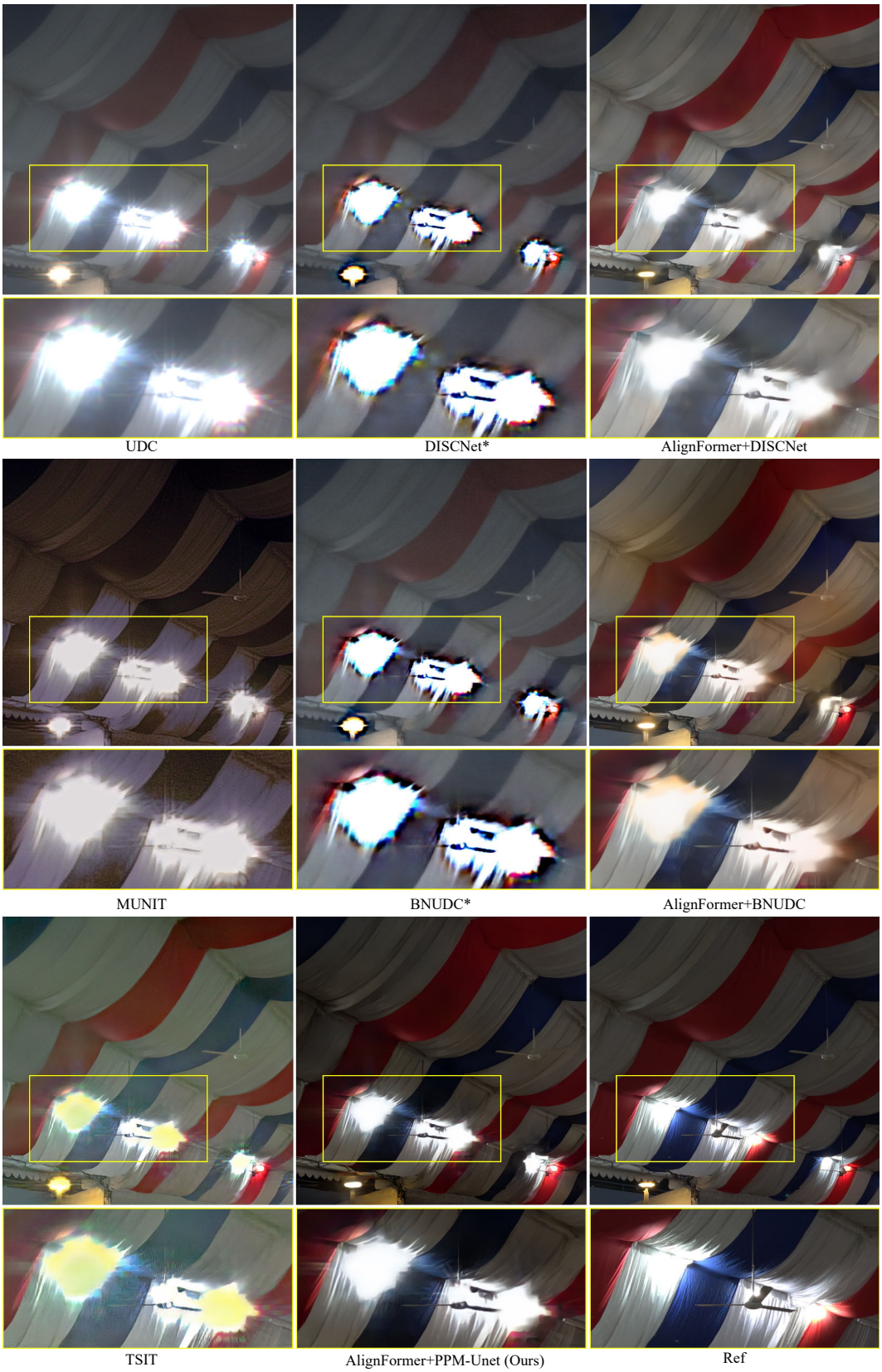


Figure A12. Qualitative comparisons on representative real-world samples.

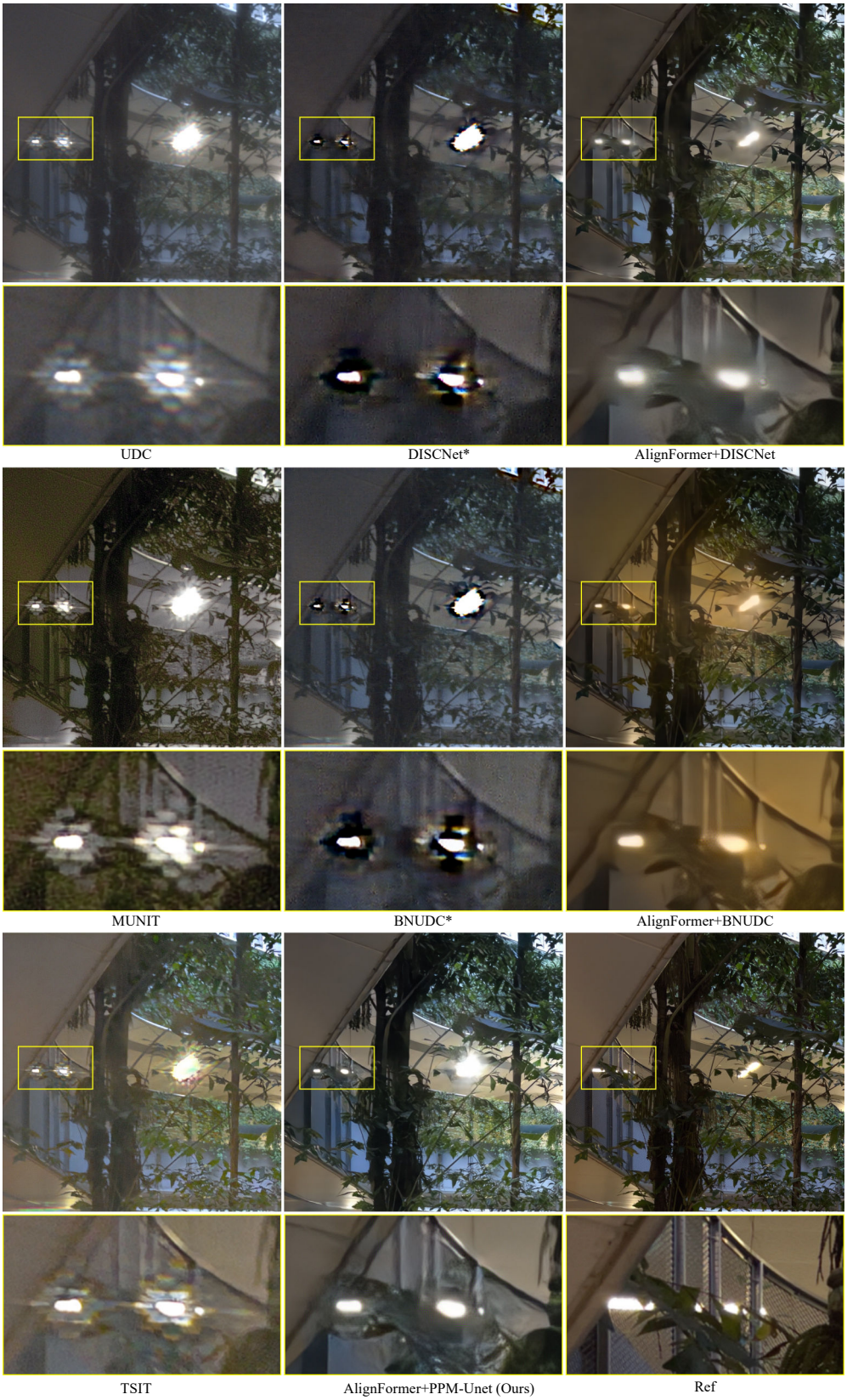


Figure A13. Qualitative comparisons on representative real-world samples.