Learning Rank-1 Diffractive Optics for Single-shot High Dynamic Range Imaging

Qilin Sun[1] Ethan Tseng[2] Qiang Fu[1] Wolfgang Heidrich[1] Felix Heide[2]
[1]KAUST [2]Princeton University

In this supplement we present additional details and results for the methods presented in the main text. Specifically, we present

- Regularization (Section 1)
- Continued optimization without rank-1 height map factorization and without regularization (Section 2)
- Scene depth experiments (Section 3)
- Limitations (Section 4)
- Details for reconstruction network (Section 5)
- Training details (Section 6)
- Details on ablation study (Section 7)
- Fabrication details (Section 8)
- PSF calibration (Section 9)
- Additional results (Section 10)
- Experimental setup (Section 11)
- Dataset sources (Section 12)

1. Regularization

As described in Section 3 of the paper, we apply a regularization loss during training. This loss is applied by using the energy distribution mask shown in Figure 1c and keeping 94% of the energy in the center and 6% in the line-like satellite regions. Our regularizer is formally given by

$$\mathcal{L}_{\text{reg}} = \tau_{\mathbf{c}}|0.94 - \mathbf{p} \odot \mathbf{M_c}| + \tau_{\mathbf{s}}|0.06 - \mathbf{p} \odot \mathbf{M_s}| \quad (1)$$

where $\mathbf{p}$ is the PSF, $\mathbf{M_c}$ is the energy mask corresponding to the center, and $\mathbf{M_s}$ is the energy mask corresponding to the satellite regions. In our experiments we used $\tau_{\mathbf{c}} = 0.05$ and $\tau_{\mathbf{s}} = 0.1$.

We also performed an experiment where we trained using our optical model but without regularization. We found that the final PSF converges to a Dirac point instead of spreading out energy from the saturated area and the performance is only 36.9 dB PSNR and 61.45 points on HDR-VDP 2 [2] on the test set. This experiment illustrates the importance of our regularizer.

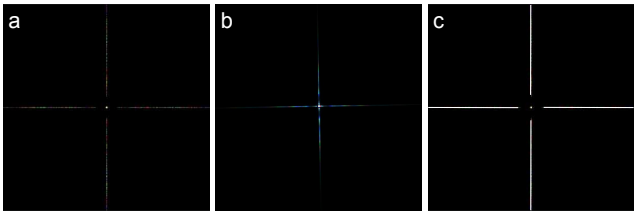

Figure 1. PSF corresponding to Rank-1 height map parameterization. (a) Simulated PSF. (b) PSF captured in the real-world. (c) Energy distribution mask used for regularization.

2. Continued optimization without rank-1 height map factorization and without regularization

To validate that we achieve a good local optimum with our optical design we continued to train without our rank-1 factorization and without our regularizer for 25 epochs. That is, we take our learned height map profile and continue to train as an unconstrained height map where each location is a learnable parameter. For this experiment the starting learning rate of the optical model is lowered from $1\text{e}{-3}$ to $1\text{e}{-6}$ while all other hyperparameters are the same as the original model training process. We observe that after 25 epochs the height map has changed insignificantly, as illustrated in Figure 2. This suggests that we do indeed find a good local optimum.

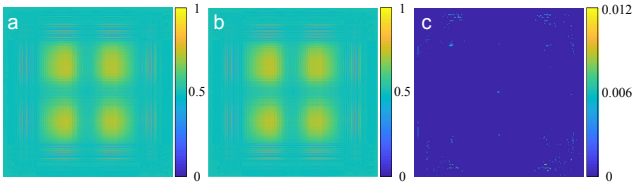

Figure 2. Height map comparison. (a) Our learned height map. (b) Continued training height map without rank-1 factorization and without regularization. (c) Absolute difference between (a) and (b). All figures are normalized by the maximum fabrication height $h_{\text{max}} = 1.125\ \mu\text{m}$

3. Scene depth experiments

Our optical model assumes that the point light source is placed 5 m away from the DOE plane. However, the PSF varies with different scene depth. As such, we investigate the robustness of our reconstruction network for handling PSFs corresponding to different scene depths in simulation. We change the position of the point light source from 1 m to infinity (while adjusting the distance between the sensor and focusing lens accordingly). Figure 3 shows our PSNR results on the test set in simulation. Our reconstruction network does best for 5 m depth as expected, and the performance is slightly degraded for other depths.

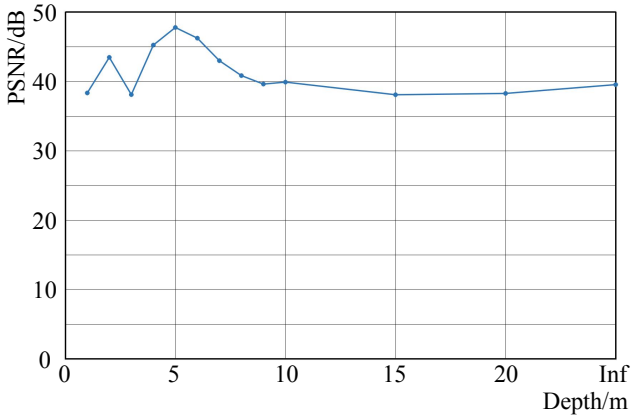

Figure 3. PSNR performance in simulation over different depths.

4. Limitations

Large saturated regions Our method is most effective for recovering highly saturated, small area regions, but struggles like other optical encoding methods when the saturated regions are larger in area.

We performed simulation experiments on scenes with larger saturated regions to further illustrate the capabilities and limitations of our method, which can be seen in Fig. 4. The left images contain large saturated regions which causes some of the encoding streaks to be saturated. In spite of this, our method is still able to recover lost details and remove the encoding artifacts. The middle images show that our method is able to accurately detect and recover highlights of different intensities even if they all lie within the same saturated region. Specifically, the high intensity ceiling lights are correctly determined to be of higher intensity than the reflected light from the windows, even though both are saturated in the LDR measurement. The right image illustrates a failure mode consisting of a very complex scene within a large saturated region. Nevertheless, our method is still able to recover details and with correct intensity levels.

Layer	Convolution layer	Activation	Normalization
0	conv-n64-k1-d1	Leaky Relu	nm
1	conv-n64-k3-d1	Leaky Relu	nm
2	conv-n64-k3-d2	Leaky Relu	nm
3	conv-n64-k3-d4	Leaky Relu	nm
4	conv-n64-k3-d8	Leaky Relu	nm
5	conv-n64-k3-d16	Leaky Relu	nm
6	conv-n64-k3-d32	Leaky Relu	nm
7	conv-n64-k3-d64	Leaky Relu	nm
8	conv-n64-k3-d1	Leaky Relu	nm
9	conv-n6-k1-d1	-	-

Table 1. Configuration of residual splitting network. In the table, "conv-n(a)-k(b)-d(c)" represents a convolution layer with a output channels, using a $b \times b$ kernel, and using a dilation rate c. Each "Leaky Relu" has slope 0.2 and $\text{nm}(x) = w_0 x + w_1 \text{Instance\_norm}(x)$, where w_0 and w_1 are trainable variables.

5. Details for reconstruction network

In Fig 5 we show intermediate network outputs for the first scene in Fig. 4 of the paper.

Figure 5. Intermediate images of our network for the first scene in Fig. 4 of the paper. $\hat{\mathbf{I}}_{\mathbb{U}}$ shows that the streak encodings have been correctly removed. $\hat{\mathbf{I}}_{\mathbb{S}}$ is displayed at -8 EV and shows accurate reconstruction of highlights. The residual image $\hat{\mathbf{I}}_r$ shows that the streaks have been identified and separated from the input image.

Details for residual splitting network Our residual splitting network configuration is shown in Table 1. As described in Section 4 of the paper, we use the pre-trained VGG-19 model to extract feature maps (1472 channels in total) and upsample them to the same size as the input image (3 channels). Then we concatenate them together (1475 channels) and feed into our residual splitting network. We use a skip connection so that the output unsaturated image estimate $\hat{\mathbf{I}}_{\mathbb{U}}$ is given by the sum of the first three channels of the output of our residual splitting network and the input image $\mathbf{I}_s$. The last three channels of the output of our residual splitting network gives the encoded residual estimate $\hat{\mathbf{I}}_r$. We clip $\hat{\mathbf{I}}_{\mathbb{U}}$ to $[0, 1]$ and $\hat{\mathbf{I}}_r$ to $[0, 2^4]$.

Details for highlight reconstruction network Our highlight reconstruction network configuration is shown in Table 2. We avoid using normalization in the last two-layers '9_1' and '9_2', and the last layer '10' to allow the network to output a larger range of values. The output of the network $\hat{\mathbf{I}}_{\mathbb{S}}$ is clipped to $[1, 2^8]$.

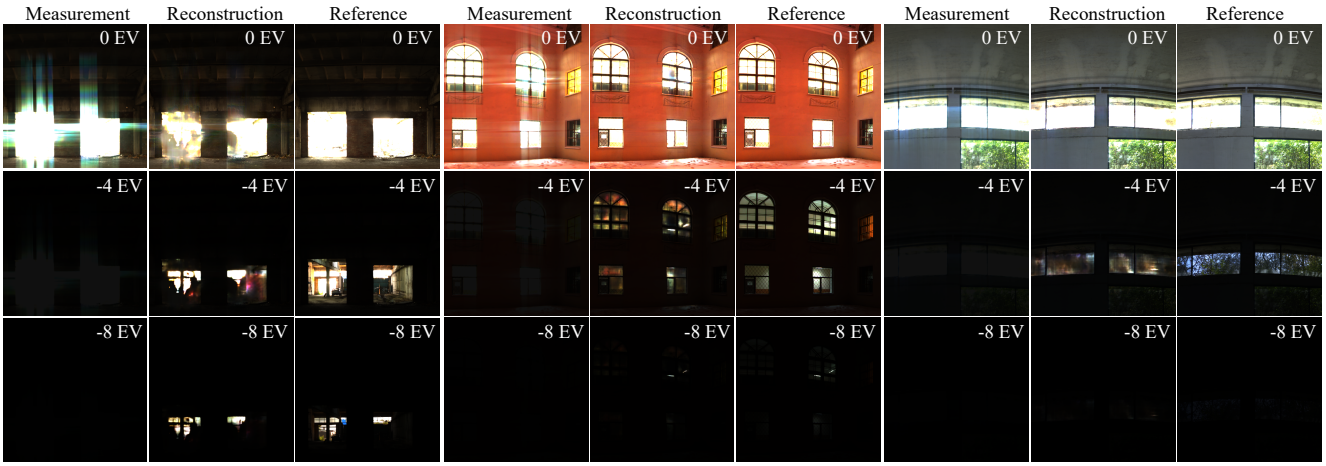

Figure 4. We provide additional simulation results on area with larger saturation regions. These larger saturated highlights are more difficult to recover, but we are still able to provide accurate reconstructions. Note that different highlight intensities within saturated regions are reconstructed with the correct intensity, for example, the ceiling lights in the middle image have higher intensity than light reflected from the windows.

Details for fusion network As shown in Table 3, we first convolve inputs $\hat{\mathbf{I}}_{\mathbb{U}}$ and $\hat{\mathbf{I}}_{\mathbb{S}}$ separately using two convolution layers for each input. Then we concatenate the feature maps together and generate the final HDR prediction using another two convolution layers. In addition, we mask out the unsaturated region for $\hat{\mathbf{I}}_{\mathbb{S}}$ before it is sent to the fusion network.

6. Training details

We implement our rank-1 DOE height map model and reconstruction network in TensorFlow 1.14. Our reconstruction network assumes inputs are in the range $[0, 1]$, and outputs are in the range $[0, 2^8]$. The model is jointly optimized using the Adam optimizer with polynomial learning rate decay. The optic was optimized using a starting learning rate of $1\text{e}{-3}$ and the reconstruction network was optimized using a starting learning rate of $1\text{e}{-4}$. The network was then fine-tuned with the trained optic by adding the exclusion loss $\mathcal{L}_{\text{excl}}$ with a starting learning rate of $1\text{e}{-5}$. Finally, we fine-tune the fusion network with a starting learning rate of $1\text{e}{-5}$. Each of the three stages is trained for 200 epochs and the learning rates are decayed after 60000 training iterations to $1\text{e}{-10}$. The model is trained on a single Titan RTX, and the training process lasts around 48 hours for every 200 epochs described above.

We used the following loss coefficients:

- $\nu_{2\_2} = 1/4.8$, $\nu_{3\_2} = 1/3.7$, $\nu_{4\_2} = 1/5.6$ for $\mathcal{L}_{\text{VGG}}$
- $\alpha_1 = 0.2$, $\alpha_2 = 0$ during training and $\alpha_1 = 0.2$, $\alpha_2 = 0.01$ during fine-tuning for $\mathcal{L}_{\mathbb{U}}$
- $\beta = 0.5$ for $\mathcal{L}_{\mathbb{S}}$

We found that using Huber loss for $\mathcal{L}_{\mathbb{F}}$ produced better results than using MAE or MSE and was better at handling the large dynamic range of the output image.

7. Details on ablation study

As illustrated in Table 1 of the paper, our method outperforms the state-of-the-art methods by more than 7 dB in PSNR and 6 points on HDR-VDP 2 [2].

HDR-CNN [1] estimates the HDR image directly from a single LDR image without using any encodings. While their method works well for small dynamic ranges, it fails to reconstruct highlights with a higher dynamic range accurately. This can be seen in Figures 10, 11, 12 where the -8 EV reconstructions show that highlight intensities are severely under-estimated.

Glare HDR [4] encodes saturated information into surrounding areas using off-the-shelf glare filters. However, their reconstruction algorithm often leaves behind strong artifacts and fails to correctly reconstruct the saturated regions. Furthermore, the algorithm takes several seconds to process a single LDR image, making it impractical for real-time applications. We also performed an experiment using our reconstruction algorithm instead of theirs while still using the "Star PSF". We found that by simply changing the reconstruction algorithm, we improve by more than 10 dB in PSNR and 11 points on HDR-VDP 2 on the test set.

Deep Optics [3] was done in parallel to our work and is most similar to our approach. For our comparison experiments, we fixed their learned "Dual Peak PSF" and only trained their reconstruction network. We found that the copied peaks produced by their PSF are easily saturated, or overlap with the saturated regions, which makes them ineffective for highlight reconstruction. We also performed an

Layer	Convolution layer	Activation	Normalization
1_1	conv-n32-k7-d1	Leaky Relu	Instance
1_2	conv-n32-k3-d1	Leaky Relu	Instance
Max Pooling			
2_1	conv-n64-k3-d1	Leaky Relu	Instance
2_2	conv-n64-k3-d1	Leaky Relu	Instance
Max Pooling			
3_1	conv-n128-k3-d1	Leaky Relu	Instance
3_2	conv-n128-k3-d1	Leaky Relu	Instance
Max Pooling			
4_1	conv-n256-k3-d1	Leaky Relu	Instance
4_2	conv-n256-k3-d1	Leaky Relu	Instance
Max Pooling			
5_1	conv-n512-k3-d1	Leaky Relu	Instance
5_2	conv-n512-k3-d1	Leaky Relu	Instance
Upsampling & Concat			
6_1	conv-n256-k3-d1	Leaky Relu	Instance
6_2	conv-n256-k3-d1	Leaky Relu	Instance
Upsampling & Concat			
7_1	conv-n128-k3-d1	Leaky Relu	Instance
7_2	conv-n128-k3-d1	Leaky Relu	Instance
Upsampling & Concat			
8_1	conv-n64-k3-d1	Leaky Relu	Instance
8_2	conv-n64-k3-d1	Leaky Relu	Instance
Upsampling & Concat			
9_1	conv-n32-k3-d1	Leaky Relu	-
9_2	conv-n32-k3-d1	Leaky Relu	-
10	conv-n3-k1-d1	-	-

Table 2. Configuration of highlight reconstruction network. In the table, "conv-n(a)-k(b)-d(c)" represents a convolution layer with a output channels, using a $b \times b$ kernel, and using a dilation rate c. Each "Leaky Relu" has slope 0.2 and "Max Pooling" represents a max pooling layer with a 2×2 kernel and a stride of 2. Each "Upsampling" represents nearest neighbor upsampling with a factor 2 followed by a convolution layer with a 3×3 kernel.

experiment using our reconstruction network while using their Dual Peak PSF. By changing the reconstruction network, we improve by more than 2 dB PSNR and 2 points on HDR-VDP 2 on the test set.

Finally, we observed that if we used our reconstruction network but varied the PSF between the Dirac PSF (no encoding), the Star PSF, the Dual Peak PSF, and our PSF, then using our PSF has the best performance. This suggests that our PSF provides the best encoding for our reconstruction network.

Layer	Convolution layer	Activation	Normalization
$1\_1_{\mathbb{U}}$	conv-n64-k3-d1	Leaky Relu	-
$1\_2_{\mathbb{U}}$	conv-n64-k3-d1	Leaky Relu	-
$1\_1_{\mathbb{S}}$	conv-n64-k3-d1	Leaky Relu	-
$1\_2_{\mathbb{S}}$	conv-n64-k3-d1	Leaky Relu	-
Concat			
2_1	conv-n32-k3-d1	Leaky Relu	-
2_2	conv-n3-k3-d1	Leaky Relu	-

Table 3. Configuration of fusion network. In the table, "conv-n(a)-k(b)-d(c)" represents a convolution layer with a output channels, using a $b \times b$ kernel, and using a dilation rate c. Each "Leaky Relu" has slope 0.2 and $\text{nm}(x) = w_0 x + w_1 \text{Instance\_norm}(x)$, where w_0 and w_1 are trainable variables. $1\_1_{\mathbb{U}}$, $1\_2_{\mathbb{U}}$ are applied to $\hat{\mathbf{I}}_{\mathbb{U}}$ while $1\_1_{\mathbb{S}}$, $1\_2_{\mathbb{S}}$ are applied to $\hat{\mathbf{I}}_{\mathbb{S}}$.

8. Fabrication details

The optimized DOE is fabricated by multilevel photolithography techniques on a fused silica wafer. Since it is difficult and costly to fabricate continuous height profiles in the micro-scale, we first slice the continuous height map into $N = 2^4$ levels. This allows us to approximate the continuous target shape with 16 staircases to compromise between manufacturability and diffraction efficiency because 16-level DOEs offer $> 90\%$ diffraction efficiency while providing good control over alignment between adjacent layers.

The fabrication procedure consists of two major parts, photolithography and reactive ion etching (RIE). The photolithography step is used to form and transfer desired patterns onto the substrate. The sliced binary pattern is first written by a Heidelberg DWL 2000 laser direct writer on a 5 inch soda-lime mask. Each pixel in the mask is $6\ \mu\text{m} \times 6\ \mu\text{m}$. For the substrate, we use a 4 inch fused silica wafer. It is sputter deposited with 200 nm Chrome (Cr) as a reflective layer, and then spin-coated with $0.6\ \mu$m thick photoresist AZ1505. Next, the mask and substrate are brought together through an i-line contact aligner EVG6200 for precise alignment between the two. The typical alignment error that can be achieved is $\pm 1\ \mu$m. Once alignment is done, the wafer is exposed to UV light with $15\ \text{mJ}/\text{cm}^2$ dose. The exposed wafer is then developed in AZ MIF726 developer for 20 s to generate the pattern on the photoresist. To transfer the pattern from photoresist to Cr, we use Cr etchant to remove the Cr in open areas. The photoresist is then removed by acetone. At the end of this step, we have a patterned Cr layer on a fused silica wafer.

The RIE step is then used to create final height reliefs in the substrate. We use a mixture of Sulfur tetrafluoride (SF_4) and Argon (Ar) gases at $10\ ^\circ$C as the plasma source. The etching depths are time controlled and monitored by measurement on a profilometer. In each RIE cycle, we double

the depth that is done in the previous step in order to approximate 2π phase modulation. We design the DOE for 550 nm wavelength and the etching depths are 75 nm, 150 nm, 300 nm, and 600 nm respectively. After the etching, we remove the residual Cr layer by Cr etchant.

We apply successive iterations of the photolithography and RIE steps to have the final 16-level DOE. The final dimension of the sample is 20 mm $\times$ 20 mm $\times$ 0.5 mm.

9. PSF calibration

To obtain the high dynamic range real-world PSF, we place a point white light source 5 m away from the sensor. We take three images in rapid succession at 0 EV, -4 EV, and -8 EV, which we then combine into one HDR PSF. We then fine-tune our trained reconstruction network using the obtained HDR PSF. Our fine-tuning process lasts for 200 epochs and uses a starting learning rate of $1\mathrm{e}{-5}$ and polynomial decay after 60000 training steps to $1\mathrm{e}{-10}$.

Since the DOE is not installed on the aperture plane, the shift-invariance of the PSF is not guaranteed at every position on the sensor. Nevertheless, as shown in Figure 6 we demonstrate that the PSF is almost constant across the field-of-view of our designed frame size.

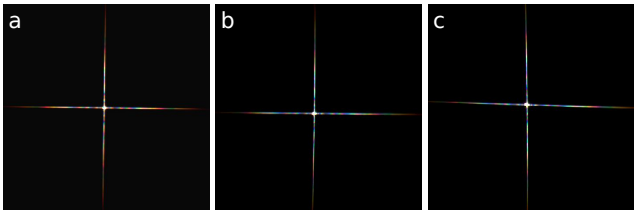

Figure 6. PSF corresponding to different sensor positions. (a) PSF located at the center of the frame. (b) PSF located at the center-left of the frame. (c) PSF located at the top-left of the frame.

10. Additional results

10.1. Real-world dynamic scene

We perform an additional experiment for capturing dynamic scenes consisting of moving high-intensity light sources. As shown in Figure 7, we capture a swinging pendulum fixture that has red and blue light sources. Since we only use one snapshot for HDR reconstruction, our method successfully recovers highlights of this dynamic scene without motion blur artifacts. However, burst HDR, which takes five images for every two stops, fails to handle this dynamic scene.

10.2. Real captures

Figures 8 and 9 show real capture results using our prototype. Note that all real captured results are shown at 1024×1024 resolution, please zoom in to see the encoding streaks and the reconstructed details.

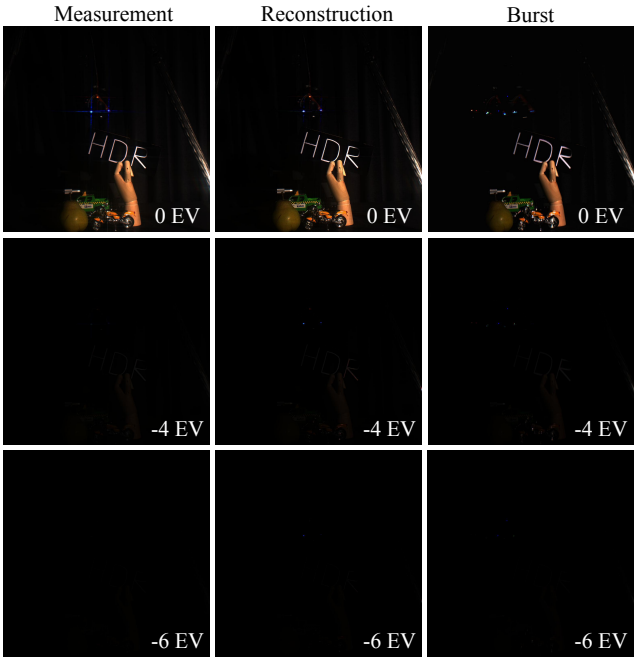

Figure 7. Visual comparison of our method against burst HDR on a dynamic scene. The scene consists of a set of lights affixed to a swinging pendulum.

10.3. Simulation comparisons

Figures 10, 11, 12 show additional qualitative comparisons in simulation.

10.4. Automotive streak removal

We model streaks using a 2-point star PSF with the same parameterization from Rouf et al. [4]. We set $\alpha = 1.0$, $\beta = 0.00025$, $\gamma = 0$, $m = 0.014$ in order to closely approximate the streaks seen in the video sequence. To remove the streaks we train our residual splitting network with the unsaturated loss $\mathcal{L}_{\mathbb{U}}$ described in Section 4.1 and use $\hat{\mathbf{I}}_{\mathbb{U}}$ as the output. We do not use the highlight reconstruction network or the fusion network for this task. Figure 13 shows additional qualitative results for automotive streak removal.

10.5. Automotive highlight reconstruction

Highlight reconstruction can also be performed with the automotive streaks. Figures 14 and 15 show highlight reconstruction results when training our full network on the same glare streaks described in Section 10.4.

11. Experimental setup

Figure 16 shows a close up frontal view of our camera prototype. Figure 17 shows a close up of our manufactured optic.

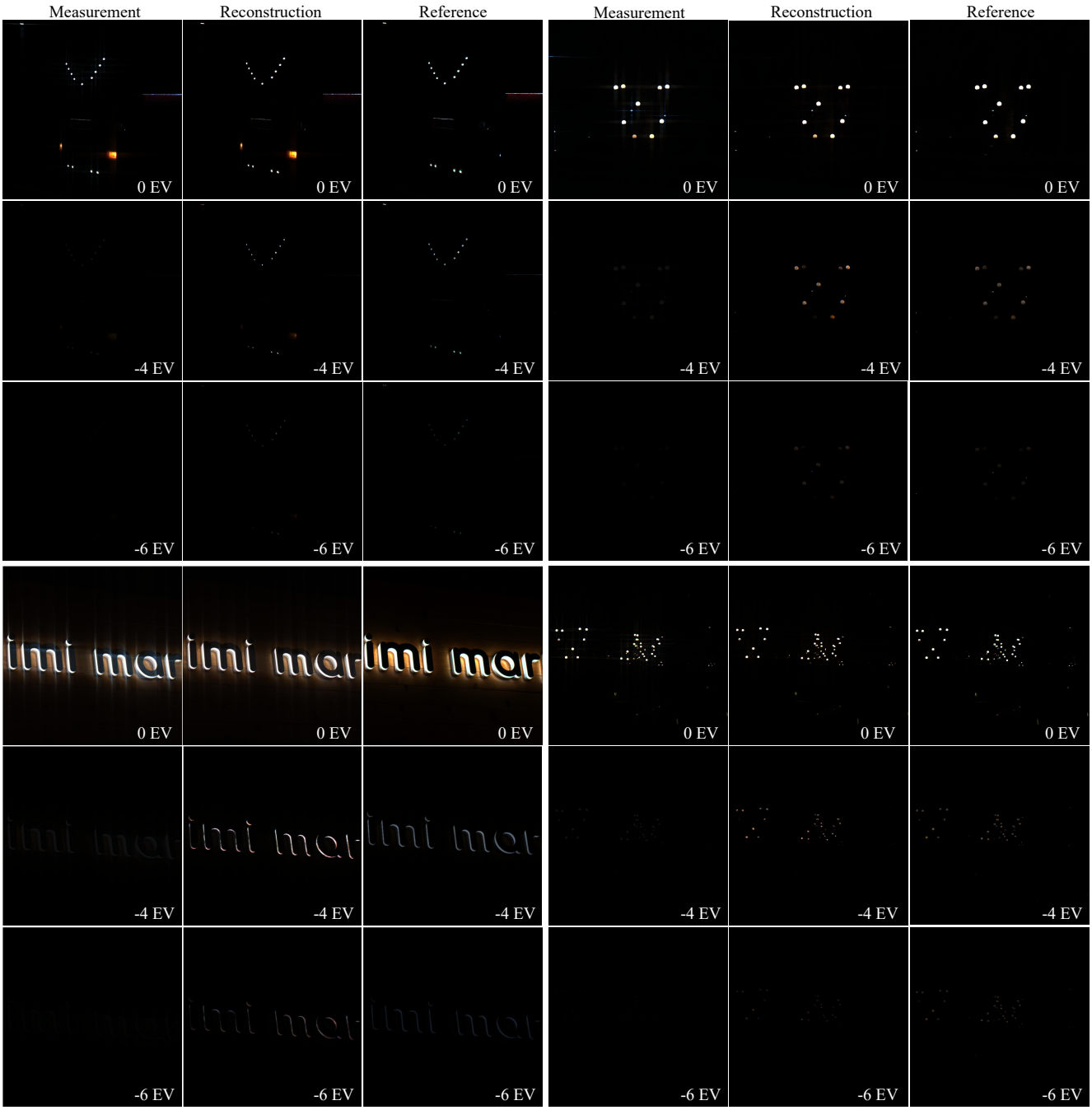

Figure 8. Additional real-world captures using our fabricated DOE prototype. Note that in the top left image set, the orange lights were shut off when taking the reference image. Please zoom in to see details.

12. Dataset sources

Table 4 shows the list of dataset sources that were used for training and testing. To accommodate different image sizes, 512×512 crops containing saturated highlights were taken.

References

[1] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017. 3

[2] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality

Figure 9. Additional real-world captures using our fabricated DOE prototype. These were earlier captures that were taken without the reference images. Please zoom in to see details.

predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):40, 2011. 1, 3

[3] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein. Deep optics for single-shot high-dynamic-range imaging. *arXiv preprint arXiv:1908.00620*, 2019. 3

[4] M. Rouf, R. Mantiuk, W. Heidrich, M. Trentacoste, and C. Lau. Glare encoding of high dynamic range images. *CVPR 2011*, pages 289–296, 2011. 3, 5

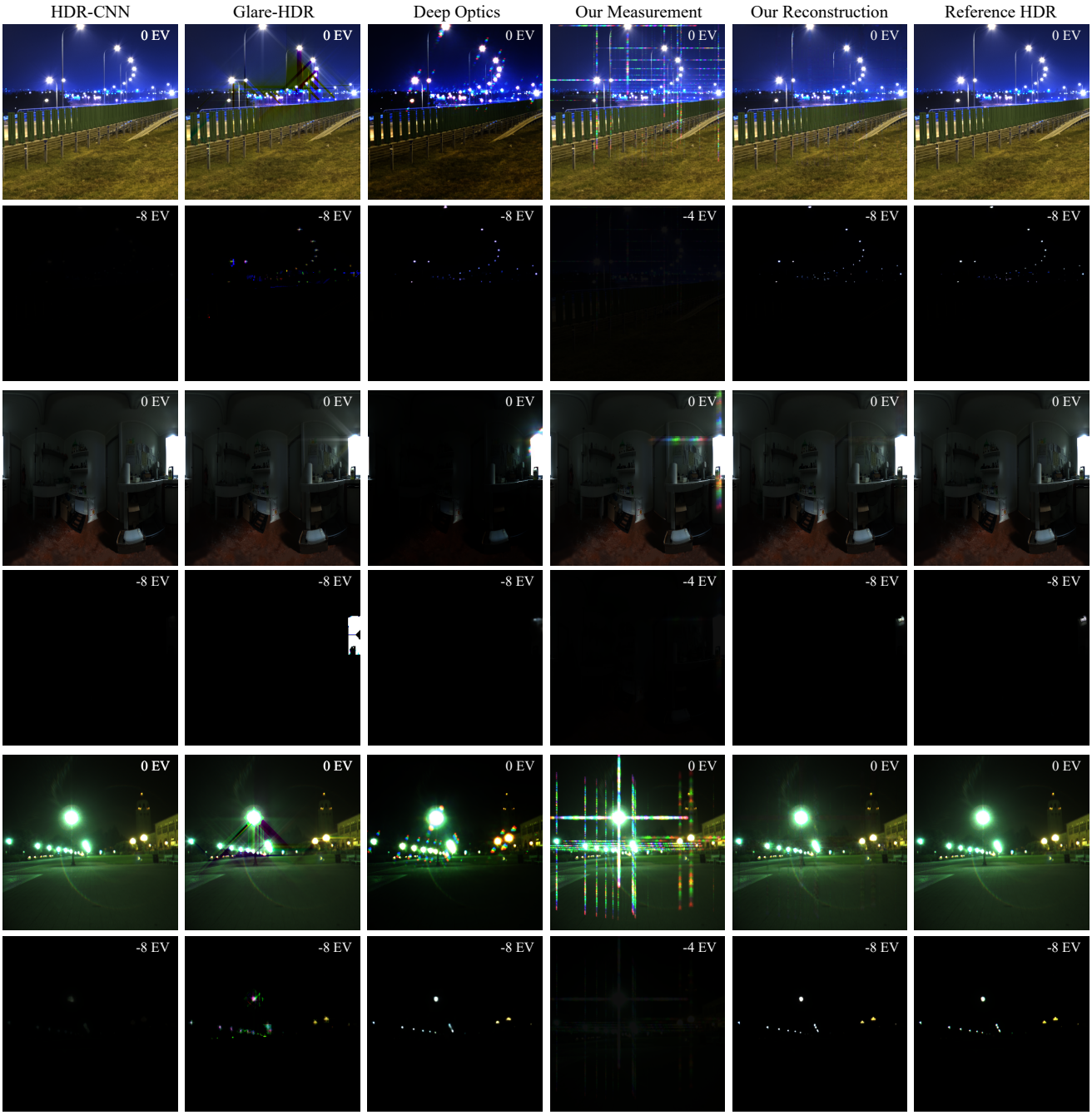

Figure 10. Additional qualitative comparisons for different snapshot HDR methods.

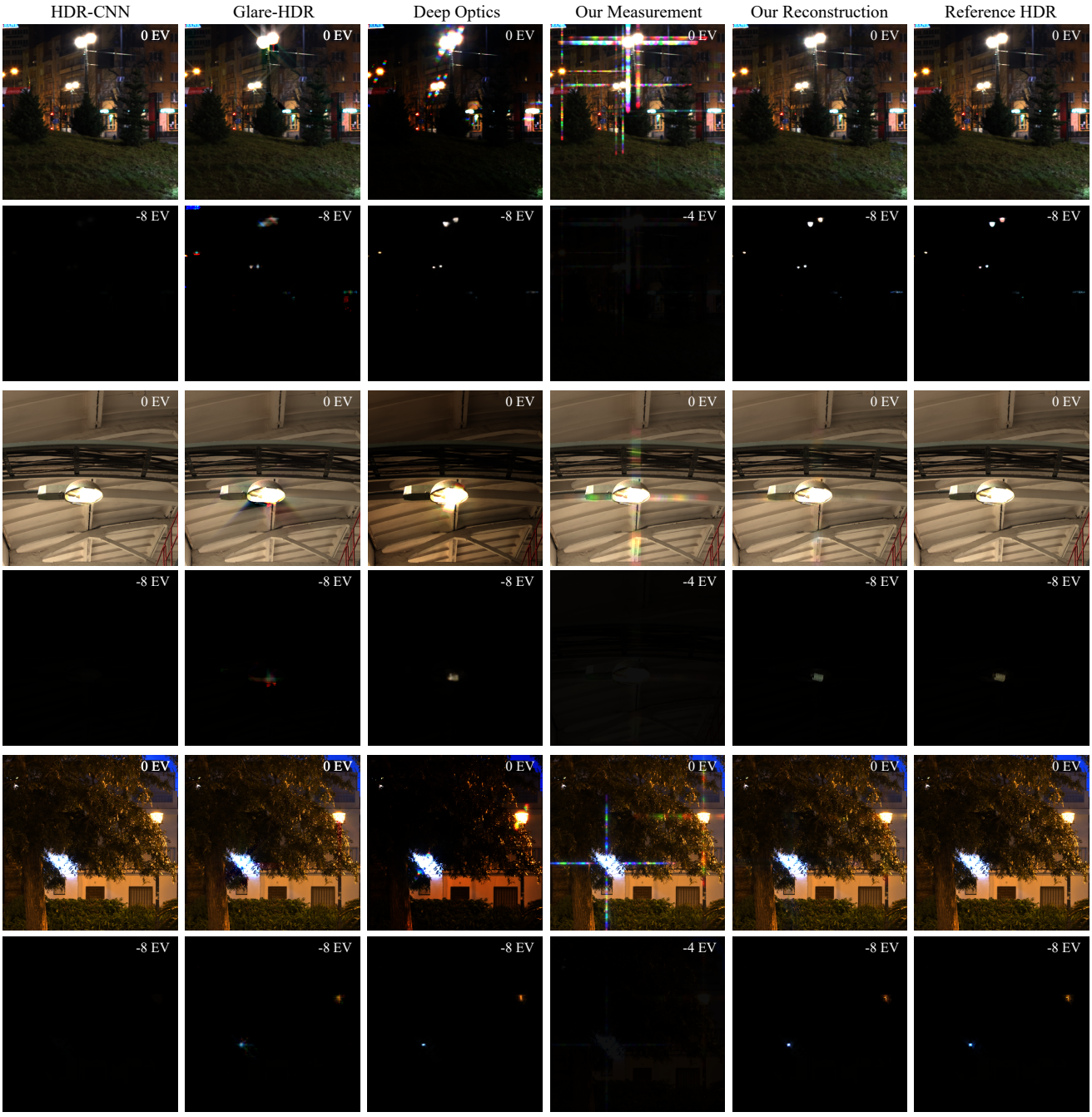

Figure 11. Additional qualitative comparisons for different snapshot HDR methods (continued).

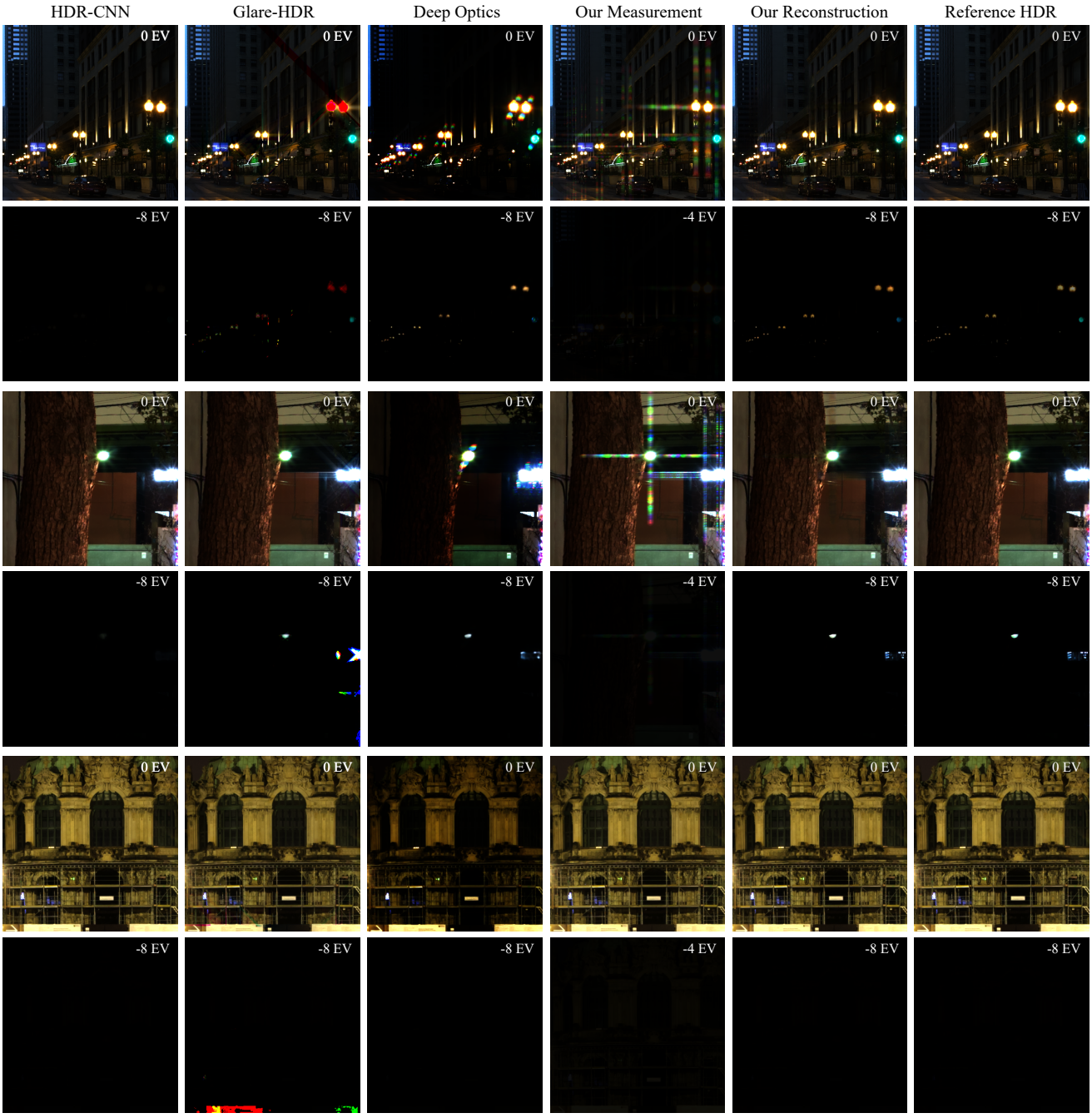

Figure 12. Additional qualitative comparisons for different snapshot HDR methods (continued). Note that the bottom image set does not contain highly saturated regions.

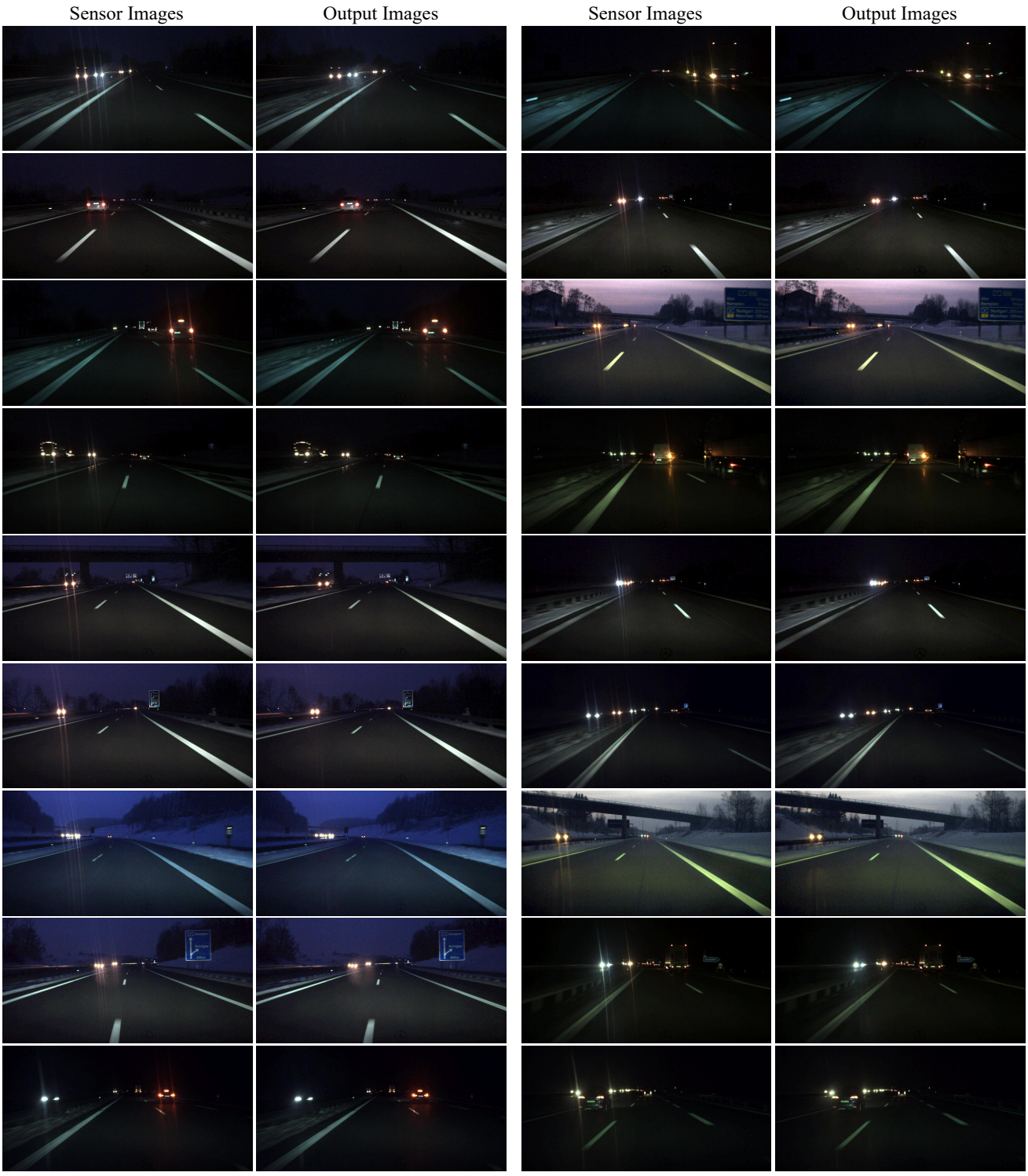

Figure 13. Additional qualitative results for automotive streak removal.

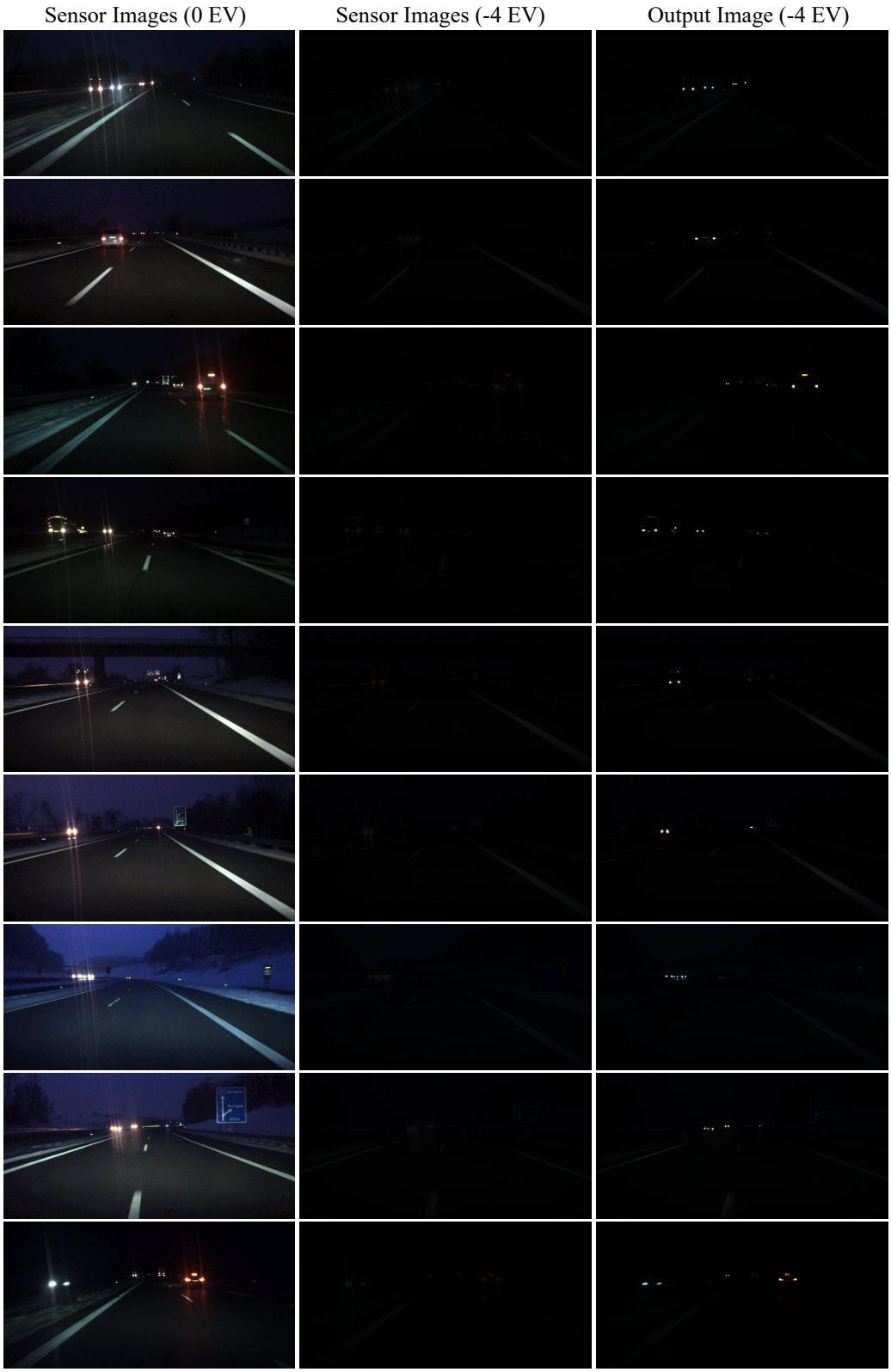

Figure 14. Qualitative results for HDR image reconstruction from automotive windshield streaks.

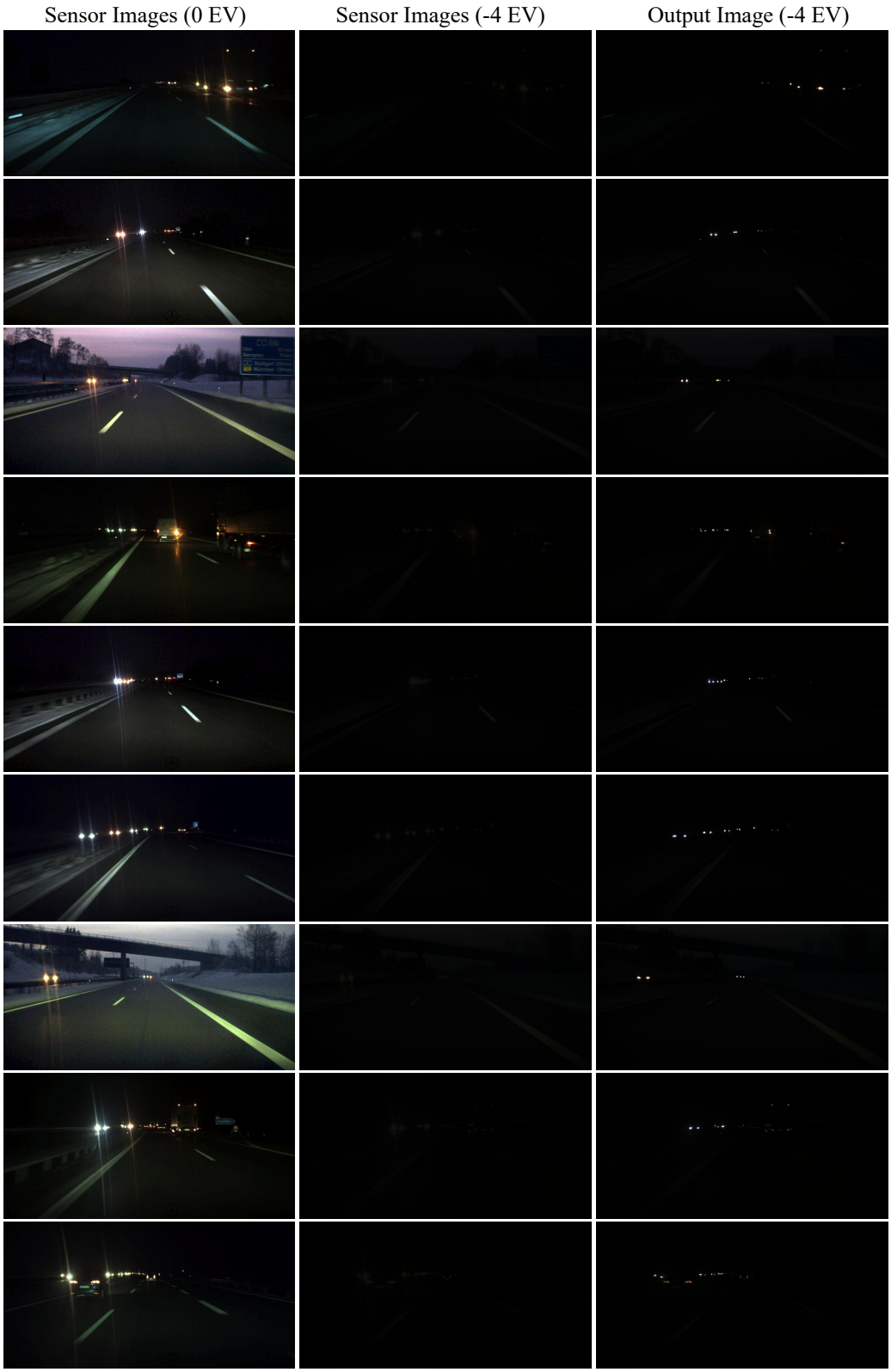

Figure 15. Qualitative results for HDR image reconstruction from automotive windshield streaks.

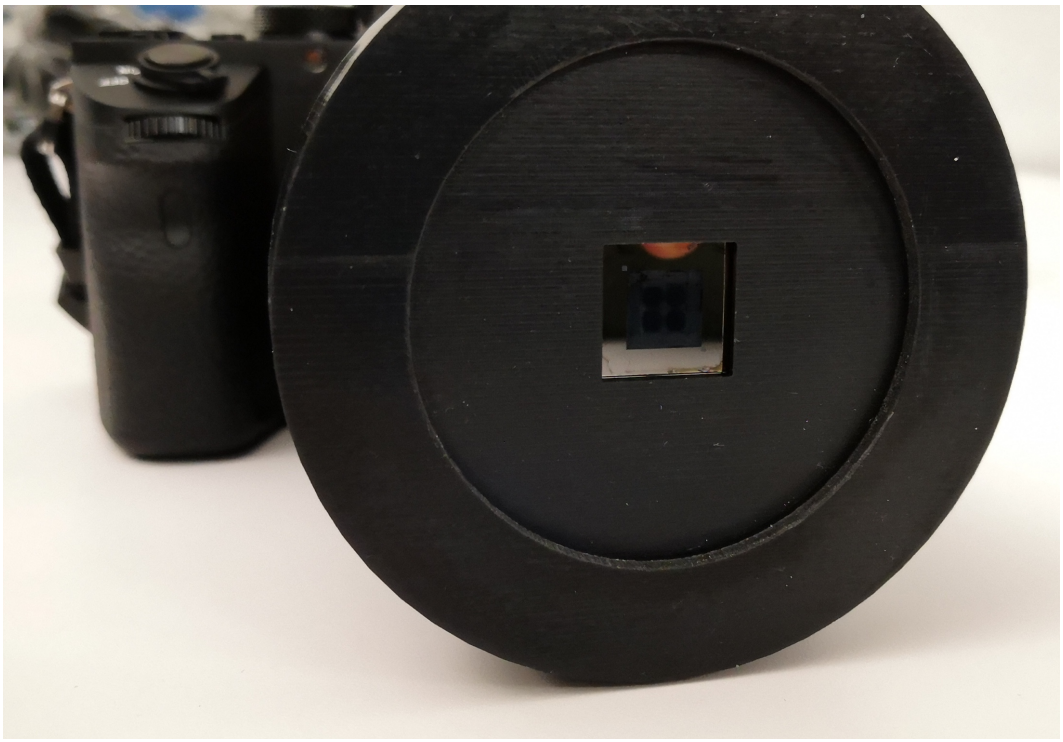

Figure 16. Close up front view of camera prototype.

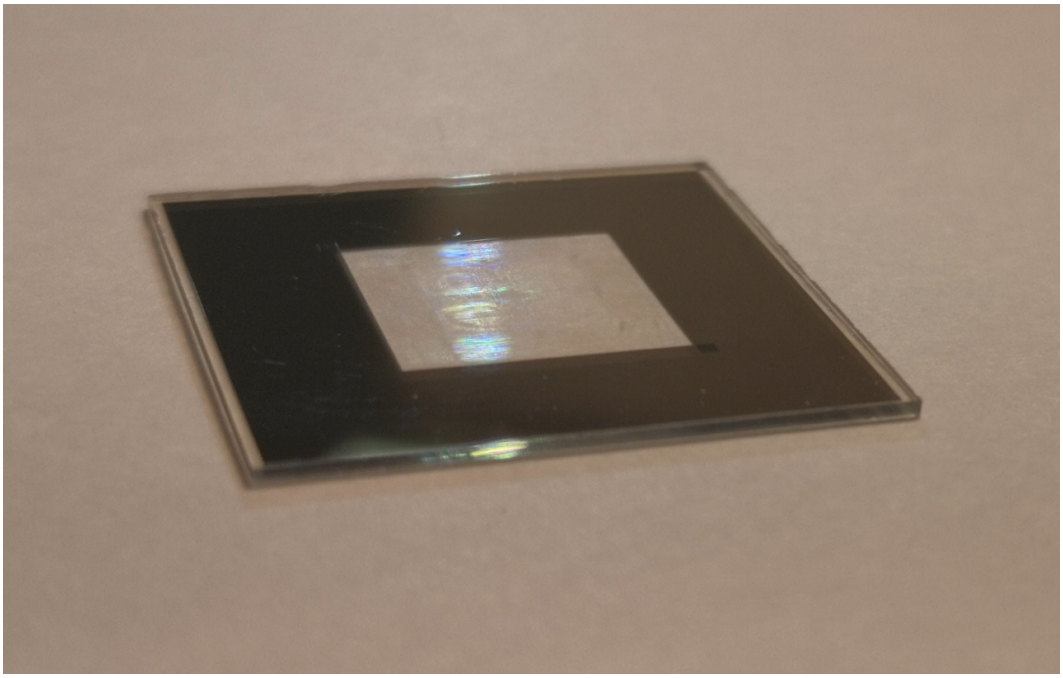

Figure 17. Close up of manufactured optic.

Name	Website	Image Names
HDRI Haven	https://hdrihaven.com	Hansaplatz, Neuer Zollhof, Preller Drive, Satara Night, Vignaioli Night, Moonless Golf, Night Bridge, Rathaus, Shanghai Bund, Zwinger Night, Rooftop Night, Carpentry Shop 1, Indoor Pool, Modern Buildings Night, Moonlit Golf, Industrial Pipe and Valve 1, Industrial Pipe and Valve 2, Viale Giuseppe Garibaldi, Winter Evening, Street Lamp, Blue Lagoon Night, Concrete Tunnel, De Balie, Garage, Boiler Room, Mutianyu, Pump House, Subway Entrance, Courtyard Night, Hospital Room, Circus Arena, Leadenhall Market, Carpentry Shop 2, Machine Shop 2, Abandoned Hall 1, Vintage Measuring Lab, Aircraft Workshop 1, Pond Bridge Night, Old Bus Depot, Entrance Hall, Small Hanger 1, Brick Lounge
HDRI Hub	https://www.hdri-hub.com	HDR City Road Night Lights, HDR Night
HiDynamic	https://shop.hidynamicproductions.com	KAH-005836-02-HDR
HDRILand	https://hdriland.com	Office Lobby, Bathroom, Sheldrake Hallway HDRI, Valley Forge Soldiers Quarters HDRI
HDRMAPS	https://hdrmaps.com	Night in Calahonda, Basketball court at night, Expressway at night, Blue hour at pier, By concert hall at night
HDRLabs	http://www.hdrlabs.com	Factory Catwalk
Joost Vanhoutte	https://joost3d.com	Amsterdam Night, Amsterdam Night 2, Amsterdam Castle, 11 Night HDRIs, 26 Free HDRIs
Ward	http://www.anyhere.com/gward/hdrenc/pages/originals.html	Atrium Night, Montreal Float
Stanford	http://scarlet.stanford.edu/~brian/hdr/hdr.html	night1, night2, night3, night4, night5, night6, night7
MCSL	Not available	Lecture Hall 2, Night car
HDRCNN	http://hdrv.org/hdrcnn/	Testset reconstructions
MPI	http://resources.mpi-inf.mpg.de/hdr/video/	Tunnel
Stuttgart	https://hdr-2014.hdm-stuttgart.de	Carousel fireworks, Beerfest lightshow
Eisklotz	https://www.eisklotz.com	Night - Church Laufenburg
LollipopShaders	http://www.lollipopshaders.com	Traffic Light on Pacifica (Night), The Parking Lot (Night)
Openfootage	https://www.openfootage.net	River power station, Trainstation Salzburg
Zwischendrin	https://www.zwischendrin.com/en/home	00065, 00080
Vlad Kuzmin	https://www.artstation.com/ssh4/store	GionSmallStreet01, SmallGion02, Gion at Night, Yard, Underpass, Tower
Corentin Defrance	https://www.artstation.com/corentindefrance	HDRI Indoor & Night Outdoor

Table 4. List of dataset sources along with specific image scenes and sets that were used.