How to cheat with metrics in single-image HDR reconstruction

Supplementary material

Gabriel Eilertsen[1], Saghi Hajisharif[1], Param Hanji[2], Apostolia Tsirikoglou[1],
Rafał K. Mantiuk[2], Jonas Unger[1]

[1] Dept. of Science and Technology, Linköping University, Sweden
[2] Dept. of Computer Science and Technology, University of Cambridge, UK

1. Dataset and camera simulation

Figure 1 shows examples of HDR scenes used in the experiments. Figure 2 demonstrates the impact of the different camera simulations that were used.

2. Complementing results

- Figure 3 complements Figure 3 in the main paper, with the same results but for EV-10 instead of EV-5.
- Figure 4 and Figure 5 complement Figure 4 in the main paper, with statistical testing of rankings for all metrics and camera simulations.
- Figure 8 complements Figure 5 in the main paper, with plots for all three different metrics.
- Figure 9 and Figure 10 complement Figure 6 in the main paper, with similar examples for other scenes.

3. Supplementary results

Figure 6 shows the same evaluations as in Figure 3 in the main paper, but evaluated only on saturated pixels. That is, if the metric is $d(\hat{H}, H)$, these results have been computed using $d(\alpha\hat{H}, \alpha H)$, where $\alpha = \max(0, L - 0.9)/0.1$ masks out only the saturated pixels. The rankings with testing of significant differences are given in Figure 7.

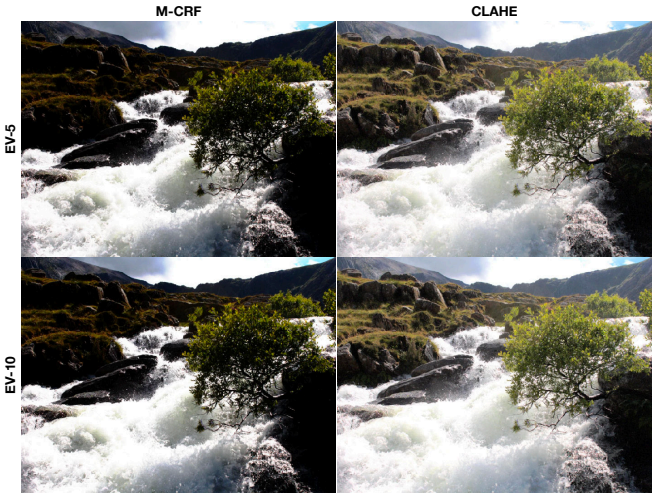
The comparison of only saturated regions can clearly separate P-rec as the best model. This is expected, since this has been composed using the ground truth information in the saturated pixels. It is also possible to see how the naive model is inferior, since this does not contain any information in the saturated pixel areas. However, for all other methods it is difficult to clearly separate between methods. Many methods also show differences in ranking depending on camera simulation.

P-lin does not contain any information in saturated regions, similar as with the naive model. Still, P-lin shows slightly higher mean. This is likely due to the blending performed by α, which incorporates also some information around the saturated regions.

The high variance, and small differences in mean between good models and the naive one, points to how this type of comparison also is inadequate, where quality of linearization still has a prominent effect on the results. There is a need for better separation of the variations in the reconstruction problem.

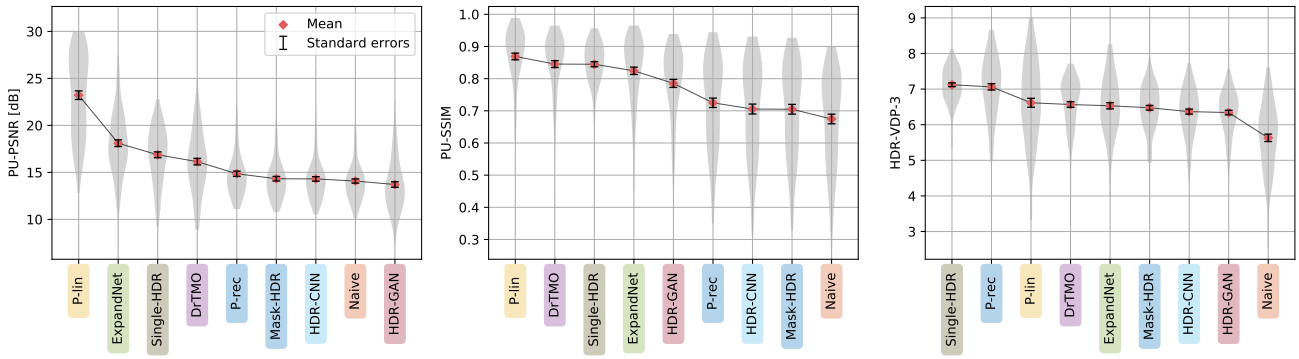
Figure 1: 15 example HDR scenes from the 96 scenes used in evaluation. Images have been gamma-encoded for display.

(a) Example 1

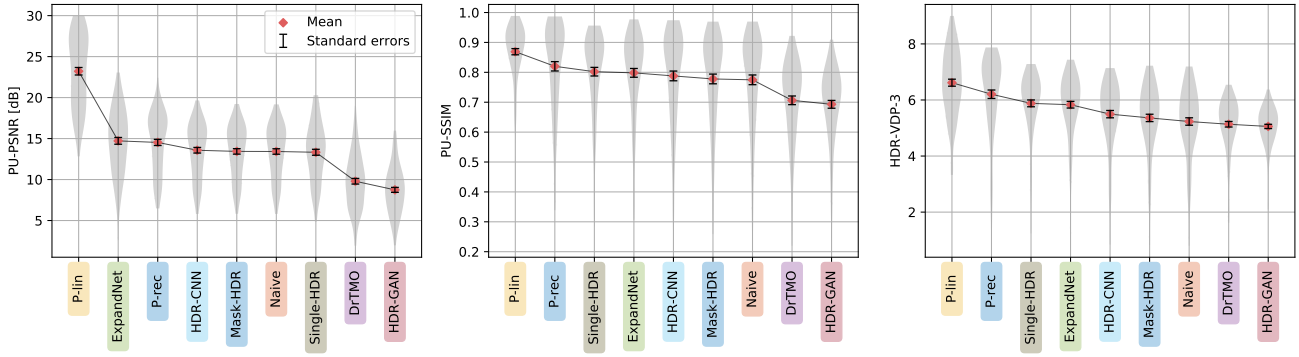


(b) Example 2

Figure 2: Examples of camera simulation, showcasing the results of different CRFs and exposures used for evaluation.



(a) Camera simulation: M-CRF, EV-10



(b) Camera simulation: CLAHE, EV-10

Figure 3: The distribution of metric values over the 96 tested scenes, where methods have been sorted by mean value to facilitate comparing differences in ranking. (a) uses camera simulation with M-CRF, while (b) is with CLAHE, and both have been simulated with EV-10. Left, middle, and right show results with PU-PSNR, PU-SSIM, and HDR-VDP-3, respectively.

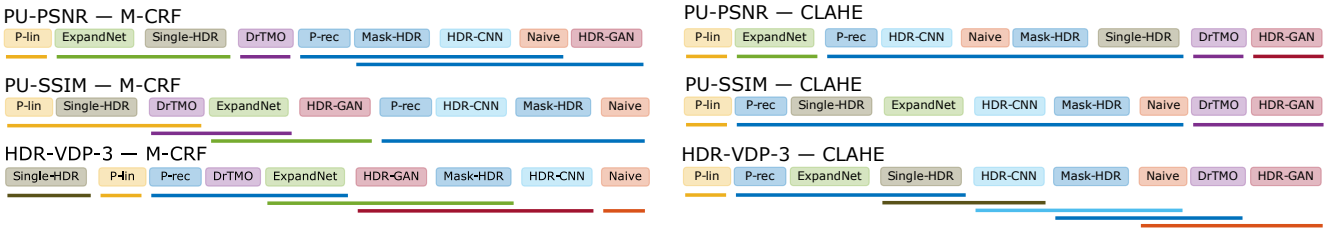


Figure 4: Rankings for different camera simulations with EV-5. The lines connect methods where the differences cannot be deemed statistically significant in a t-test, with a p-value threshold of 0.05.

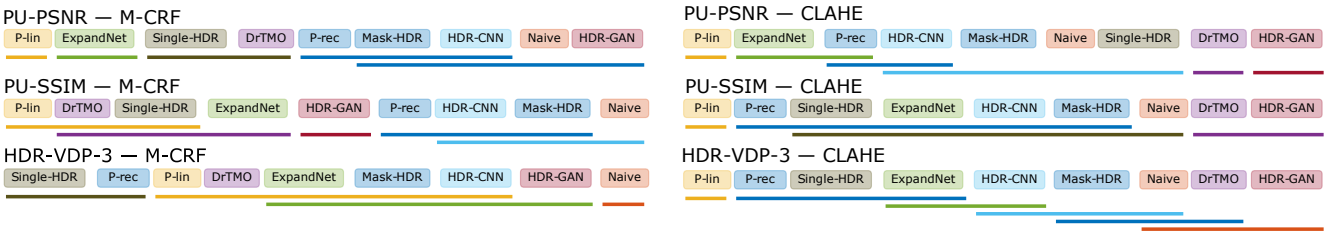
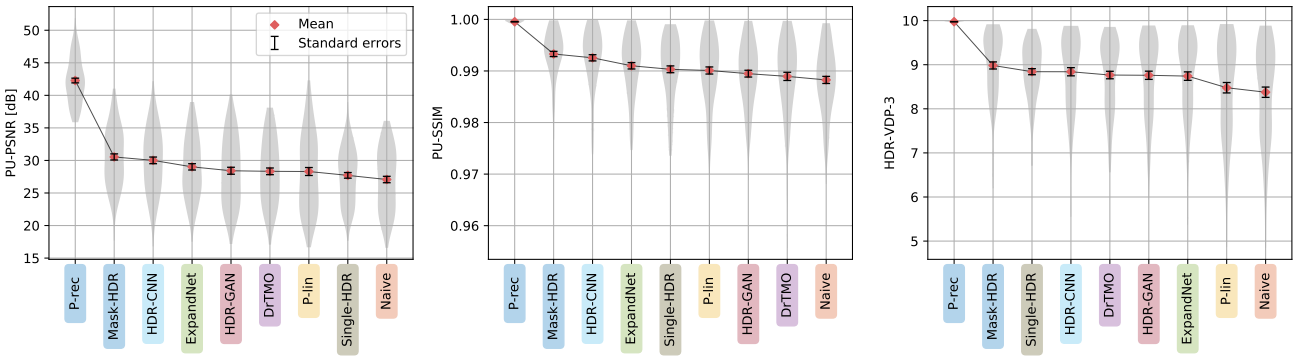
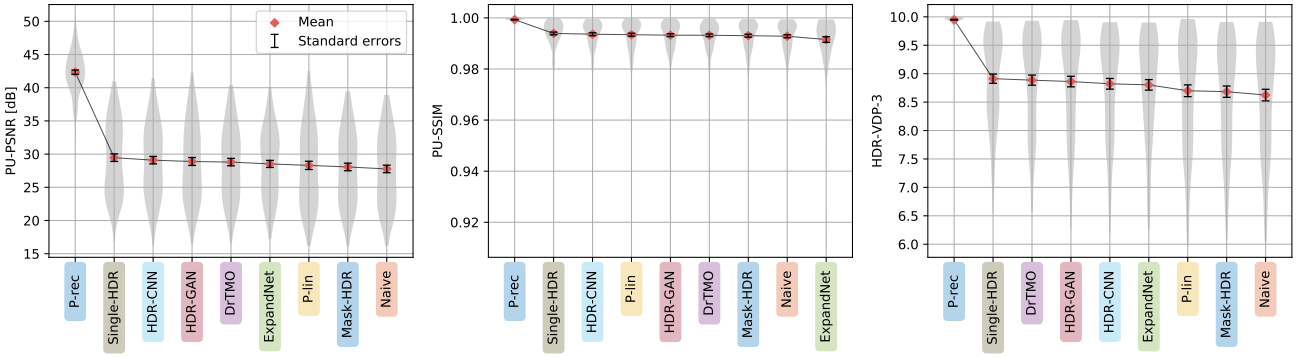


Figure 5: Rankings for different camera simulations with EV-10. The lines connect methods where the differences cannot be deemed statistically significant in a t-test, with a p-value threshold of 0.05.



(a) Camera simulation: M-CRF, EV-5



(b) Camera simulation: CLAHE, EV-5

Figure 6: Same as Figure 3, but for EV-5 and only evaluated in saturated regions of the images.

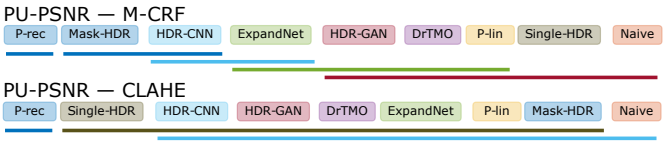
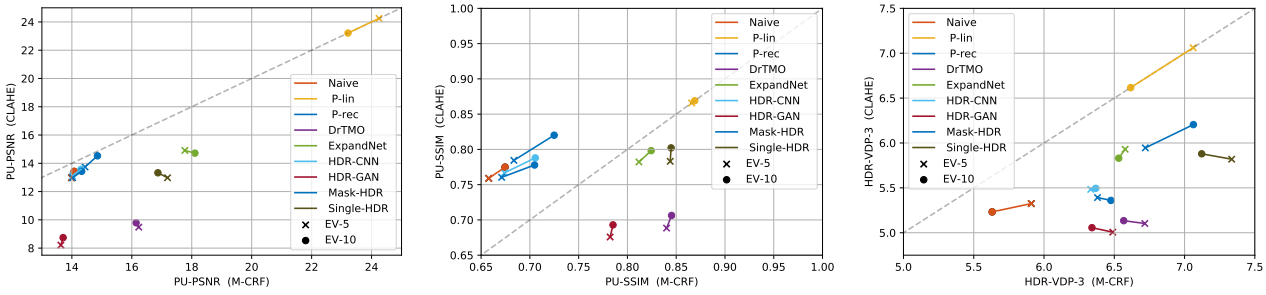
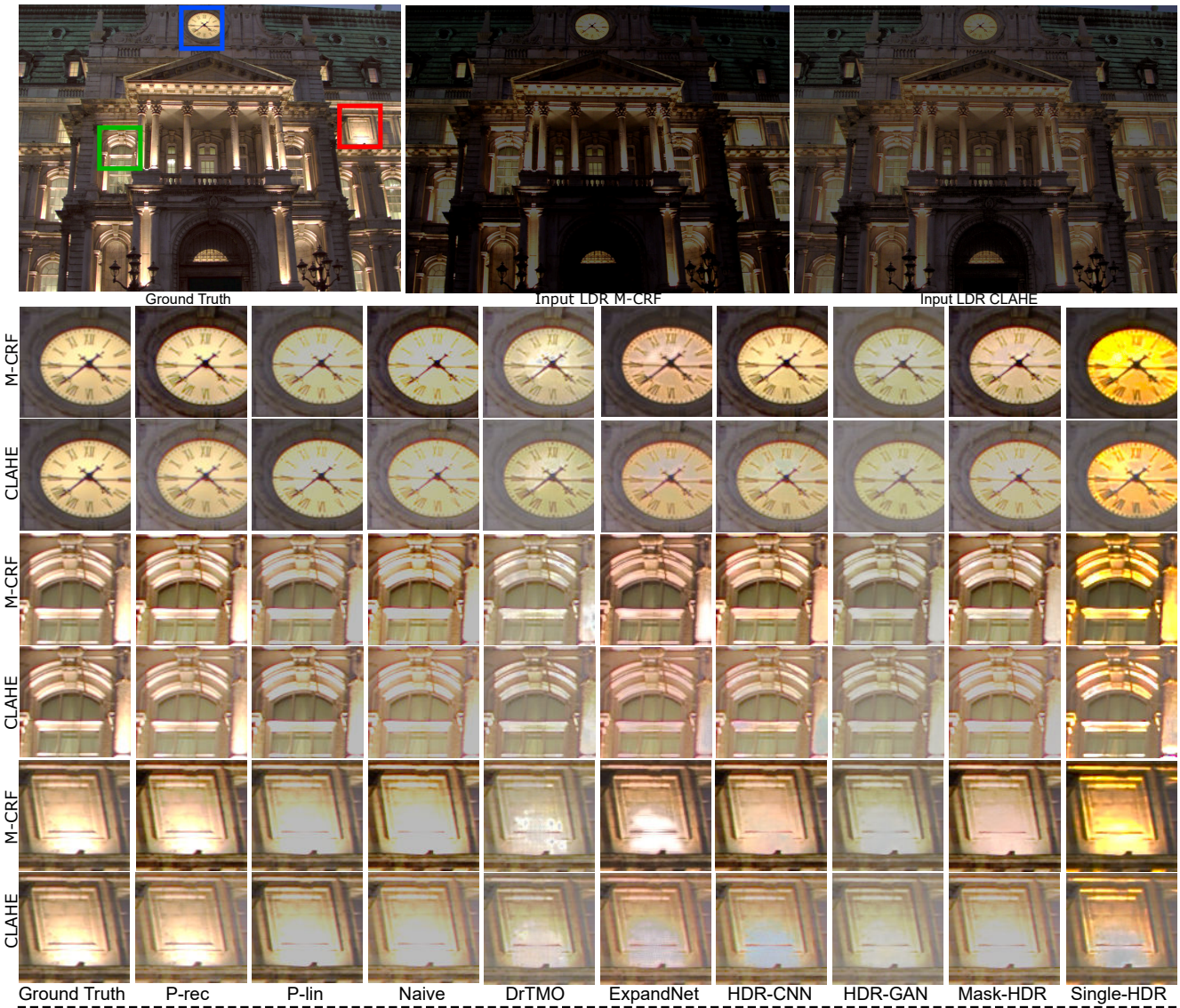


Figure 7: The rankings provided by PU-PSNR with M-CRF and CLAHE, when evaluated only in saturated image regions.



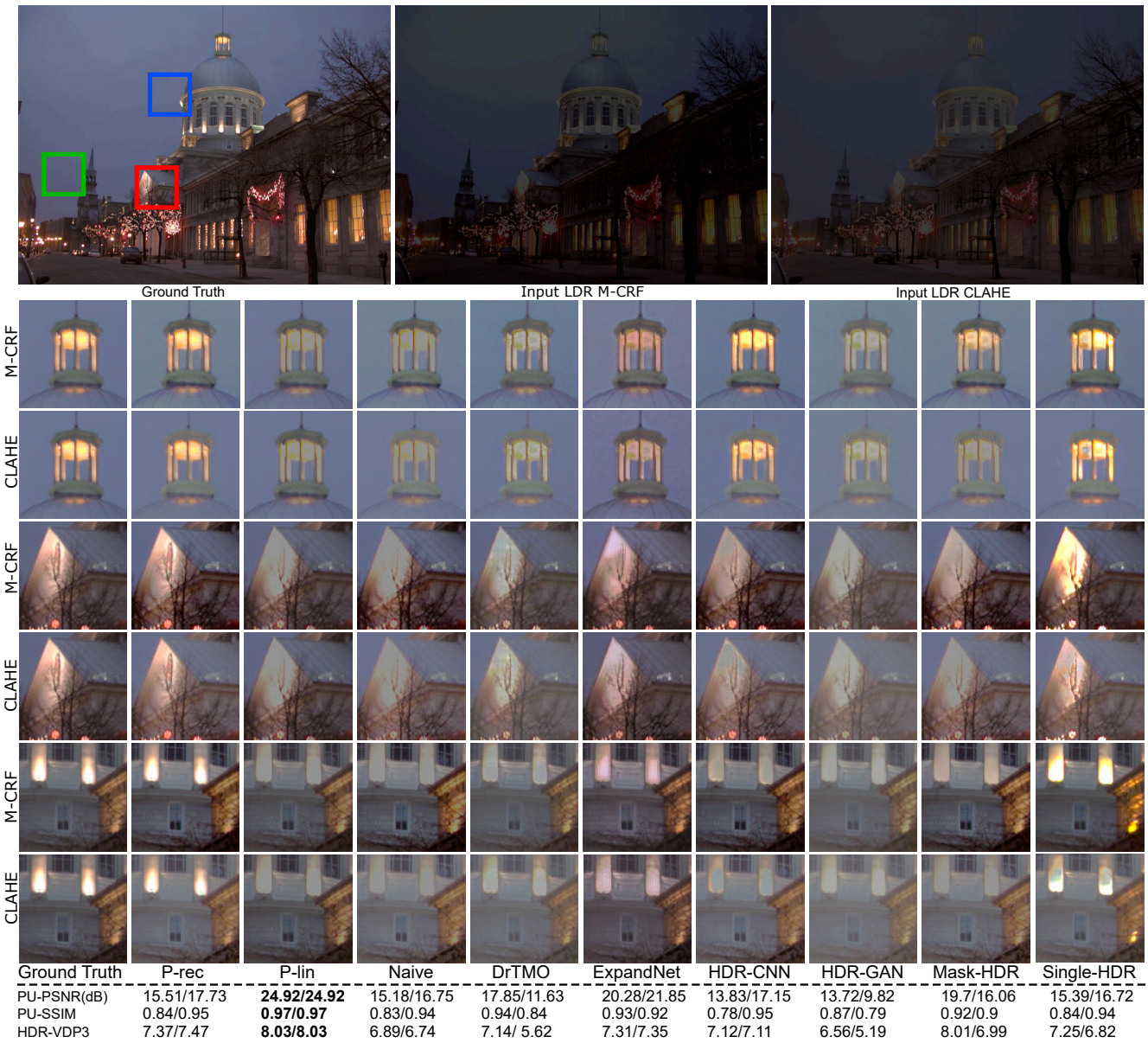
(a) PU-PSNR (b) PU-SSIM (c) HDR-VDP-3

Figure 8: Differences in PU-PNSR (a), PU-SSIM (b) and HDR-VDP-3 (c) when using M-CRF and CLAHE. The two points for each method are with EV-5 and EV-10, demonstrating how many methods do not show an expected reduction in quality with increased camera exposure (more challenging reconstruction problem).



	Ground Truth	P-rec	P-lin	Naive	DrTMO	ExpandNet	HDR-CNN	HDR-GAN	Mask-HDR	Single-HDR
PU-PSNR(dB)		12.24/13.96	**22.76/22.76**	12.01 /13.15	14.68 /8.69	16.44 /13.78	11.63 /13.56	13.47/7.56	12.23 /14.44	11.69 /13.14
PU-SSIM		0.68/0.87	**0.92/ 0.92**	0.65/0.84	0.89/ 0.71	0.86/0.88	0.67/0.86	0.82/0.69	0.8/0.88	0.66/0.84
HDR-VDP3		6.76/6.85	**7.91/7.91**	6.48/6.25	6.91/ 5.49	7.01/6.57	6.68/6.45	6.73/5.28	6.92/6.85	6.69/6.29

Figure 9: Selected scene areas for different reconstructions, with input LDR images simulated using M-CRF and CLAHE. The metrics in the bottom show the performance with M-CRF/CLAHE. The exposure time was set such that 5% of pixels are saturated (EV-5). Ground truth and reconstructed HDR images have been gamma-encoded for display.



	Ground Truth	P-rec	P-lin	Naive	DrTMO	ExpandNet	HDR-CNN	HDR-GAN	Mask-HDR	Single-HDR
PU-PSNR(dB)		15.51/17.73	**24.92/24.92**	15.18/16.75	17.85/11.63	20.28/21.85	13.83/17.15	13.72/9.82	19.7/16.06	15.39/16.72
PU-SSIM		0.84/0.95	**0.97/0.97**	0.83/0.94	0.94/0.84	0.93/0.92	0.78/0.95	0.87/0.79	0.92/0.9	0.84/0.94
HDR-VDP3		7.37/7.47	**8.03/8.03**	6.89/6.74	7.14/ 5.62	7.31/7.35	7.12/7.11	6.56/5.19	8.01/6.99	7.25/6.82

Figure 10: Selected scene areas for different reconstructions, with input LDR images simulated using M-CRF and CLAHE. The metrics in the bottom show the performance with M-CRF/CLAHE. The exposure time was set such that 5% of pixels are saturated (EV-5). Ground truth and reconstructed HDR images have been gamma-encoded for display.