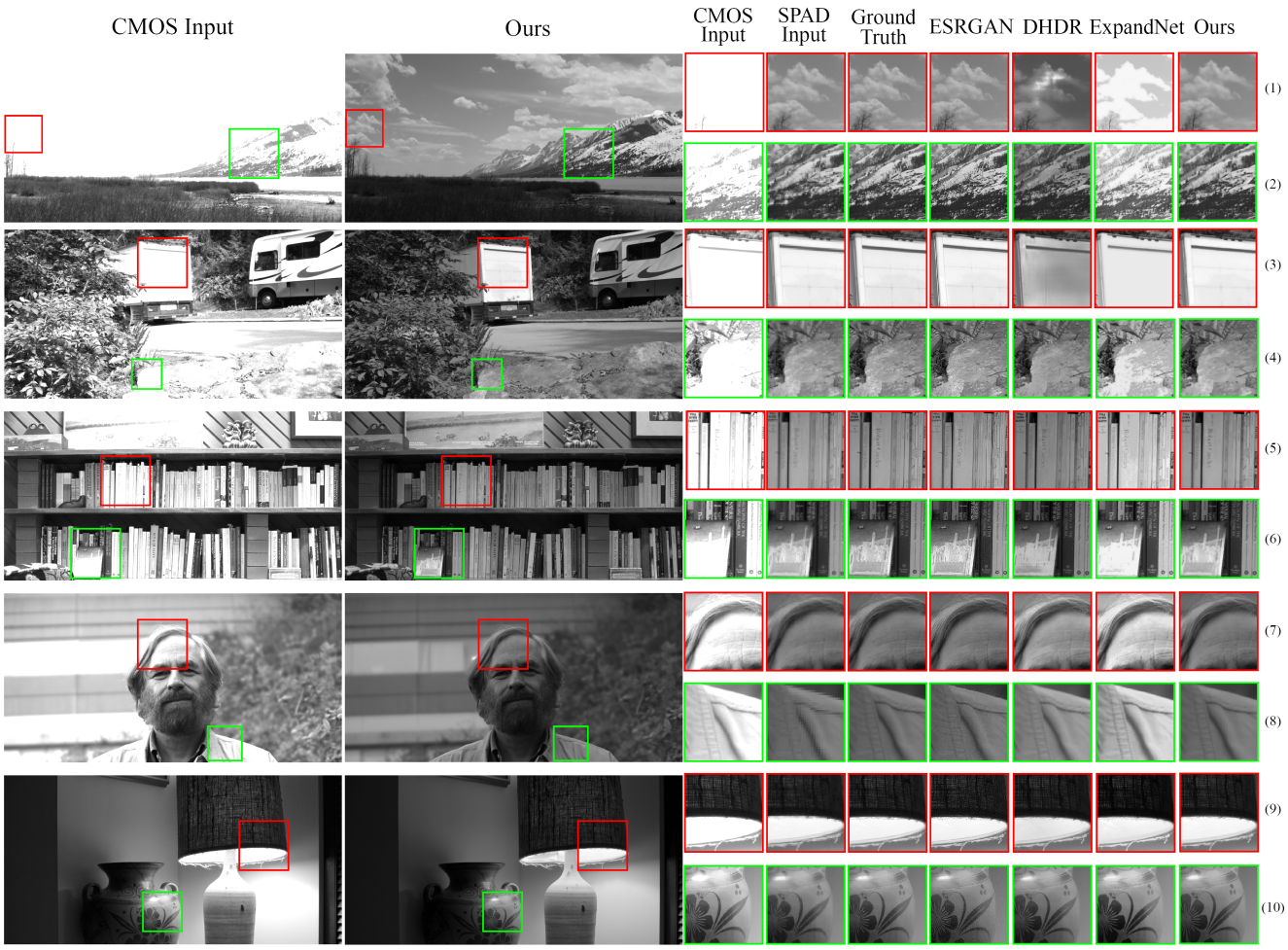
Supplementary Document for "Single-Photon Camera Guided Extreme Dynamic Range Imaging"

Yuhao Liu, Felipe Gutierrez-Barragan, Atul Ingle, Mohit Gupta, Andreas Velten

S. 1. Additional Synthetic Results

In this section we present additional qualitative and quantitative comparisons of our proposed methods against various baselines.

S. 1.1. Comparing against single-input baselines



Supplementary Figure 1. **Comparing against single-input baselines:** Visual comparison between our proposed method and single-input baselines, including ESRGAN [63], DHDR [56], and ExpandNet [44].

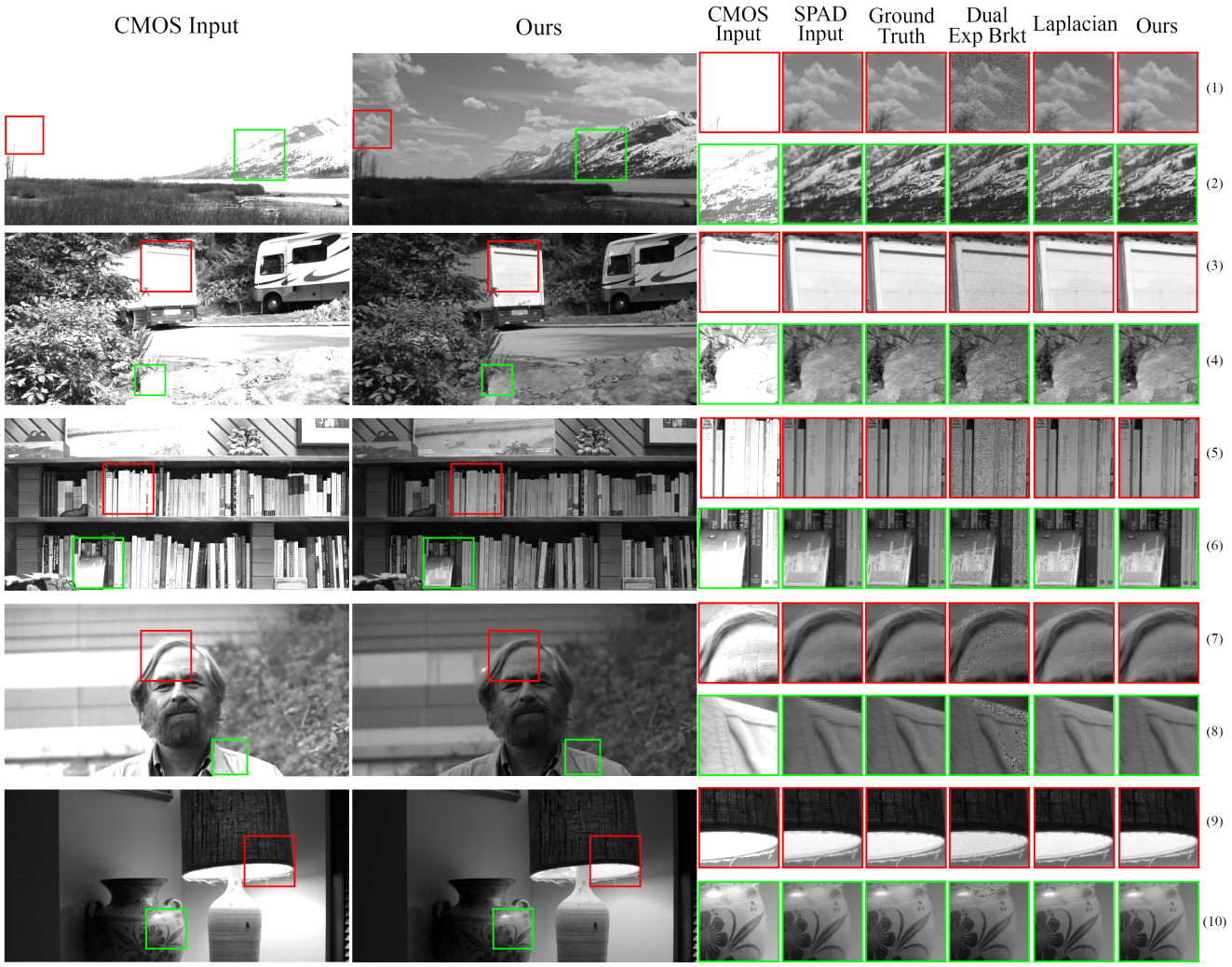
Suppl. Fig. 1 establishes a comparison between our method and all baselines that use *only* one image as input. The list below highlights the performance of each baseline:

- **ESRGAN [63]:** the network tends to generate high-frequency non-existent details, such as film grain-like pattern (row 1), hallucinated and over-sharpened textures (row 2), and non-existent lines at the center of the flower (bottom row). In particular, ESRGAN fails to recover text (row 5).
- **DHDR [56]:** in general, as the area of saturated region increases, DHDR loses its ability to generate texture details, likely due to the lack of valid contextual information in adjacent areas. For instance, DHDR fails to recover the grid lines (row 3). Moreover, since the texture details are hallucinated, the recovered appearance often deviates from ground truth, such as the shape of the cloud (row 1), surface texture of the rocks (row 4), and patterns on the metal plate (row 6).

- **ExpandNet [44]:** the regions that the network recovers are generally flat, lack texture details, and have synthetic appearances. For example, there is no texture or details on the clouds (row 1), back of the truck (row 3), etc.

Our proposed method does not have the problems found in the baselines outlined above. For instance, our method does not have high-frequency artifacts found in ESRGAN and performs well in recovering texts. Since our method uses low-resolution SPC images as guided data, our outputs closely mimic the appearances of the the objects in ground truth images, unlike the hallucinated textures in DHDR and ExpandNet.

S. 1.2. Comparing against fusion-based baselines



Supplementary Figure 2. **Additional Results: comparing against merged baselines:** Visual comparison between our proposed method and fusion-based baselines, including Exposure Bracketing [13], Laplacian [7] blending of ESRGAN (super-resolved SPC) and CMOS image. Please zoom in on the electronic version for more details.

Suppl. Fig. 2 establishes a comparison between our method and all baselines that use more than one images as input. Again, the list below highlights the performance of each baseline:

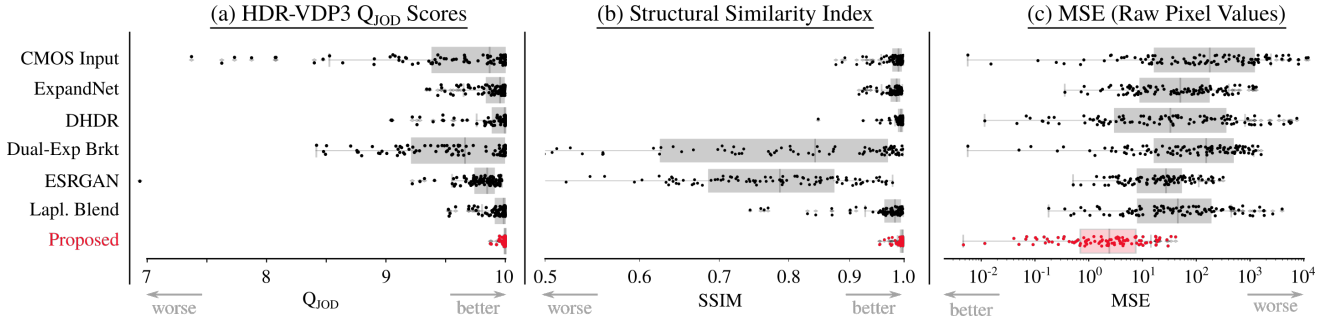
- **Exposure Bracketing [13]:** all crops contain noticeable "SNR dips", occurring at the boundary between two different exposures selected by the merging algorithm. Such SNR dips are not only visually unpleasing, but the noise also occludes image content. For example, the text information (row 5) buried under strong noise which may be very difficult to denoise.
- **Laplacian Blending [7]:** Laplacian blending fuses the CMOS image and super-resolved SPC image from ESRGAN. Visually, the Laplacian outputs look marginally better than their ESRGAN counterparts. For instance, the non-existent hallucinated lines (bottom row) are absent, and a fraction of texts (row 5) becomes readable again. However, the performance

of Laplacian is still handicapped by ESRGAN and the same issues from ESRGAN still generally apply: high-frequency artifacts (row 2), incompatability with text (row 5), etc. Additionally, the fusion boundaries are visible in some crops (row 7 and 8), which look similar to SNR dips but are significantly less severe than those found in Exposure Bracketing.

While the two baselines that we presented have apparent visible fusion boundaries, the boundaries in our method are less pronounced, since our end-to-end network fuses CMOS and SPC images without introducing high-frequency artifacts or noise, resulting in cleaner, more visually pleasing images.

S. 1.3. Quantitative Metrics

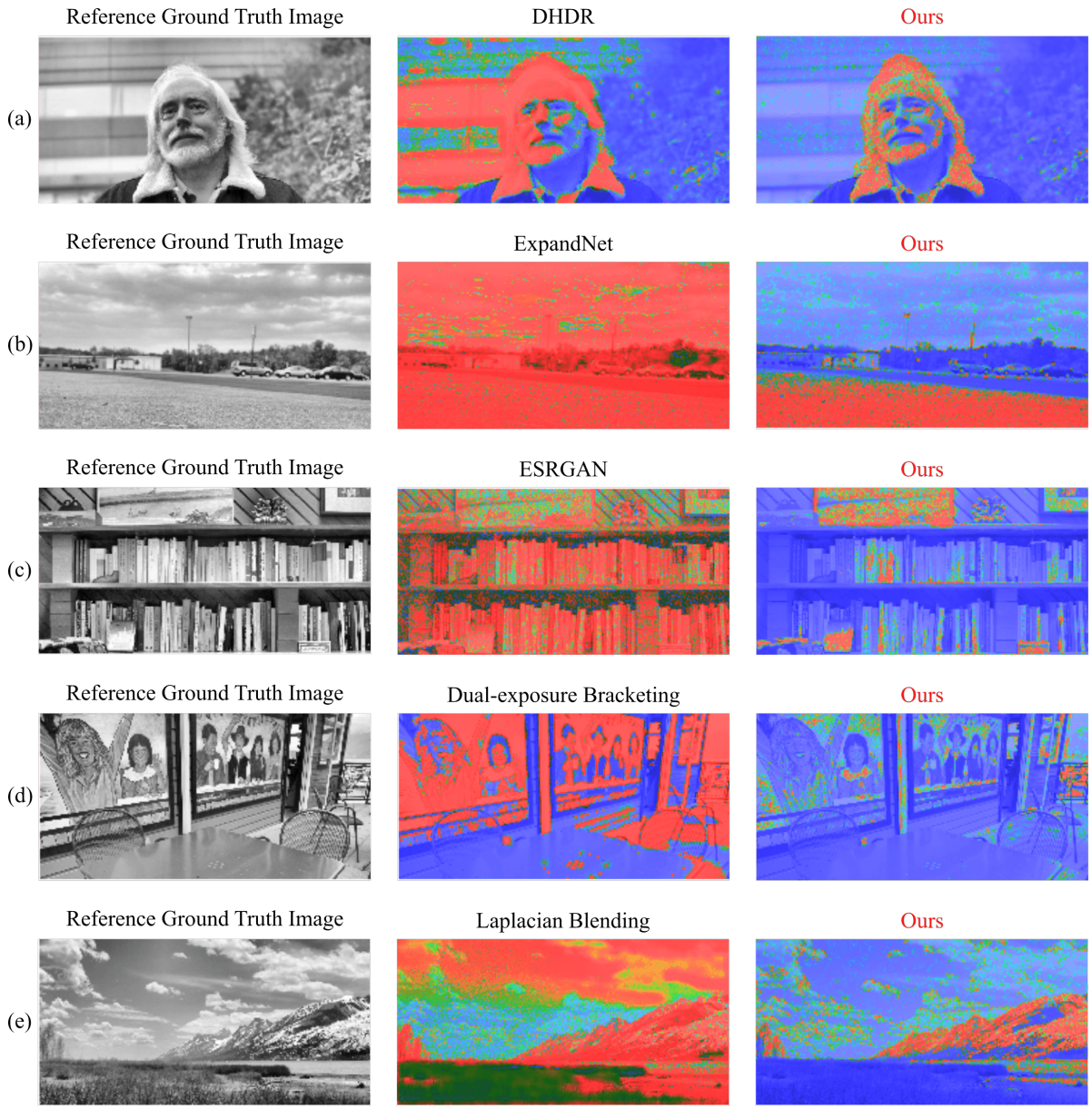
Suppl. Fig. 3 shows three quantitative metrics in addition to the two metrics shown in Fig. 7. Our method achieves higher HDR-VDP Q_{JOD} and SSIM scores than all baselines. We also achieve lower MSE than all baselines, where MSE is computed directly with the raw pixel values (without any tone-mapping or μ-compression).



Supplementary Figure 3. **Additional Quantitative Metrics for Test Set Images:** (a) HDR-VDP "just-observable-difference" quality score (Q_{JOD}), (b) structural similarity index (SSIM) computed on μ-compressed ($\mu = 500$) images, (c) raw mean-squared error (MSE) values with respect to the ground truth pixel values.

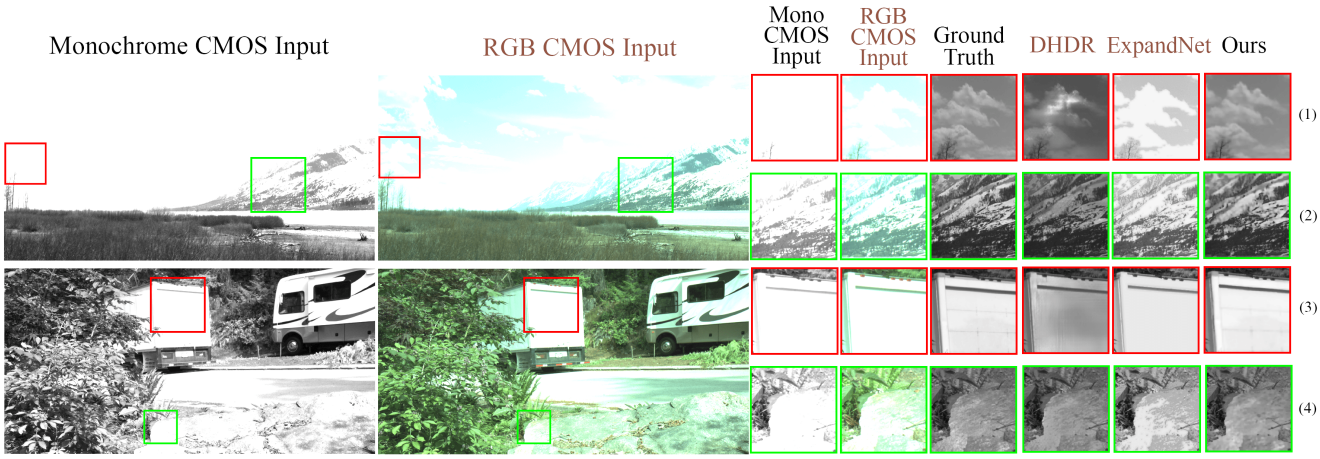
Suppl. Fig. 4 shows spatial maps of the HDR-VDP quality metric for different scenes and baselines. For instance, single-image HDR networks tend to struggle as the area of over-exposed regions become to large, regardless if the corresponding ground truth region is smooth (sky region in Suppl. Fig. 4(b)) or contains high-frequency features (white hair in Suppl. Fig. 4(a)). Single-image super-resolution methods hallucinate high-frequency textures, causing low quality scores as seen in the case of ESRGAN in Suppl. Fig. 4(c). Dual-exposure bracketing suffers from perceptible noise artifacts in the mid-photon-flux regions due to an SNR dip. This is evident in Suppl. Fig. 4(d). In Suppl. Fig. 4(e), we show the quality map for the result of Laplacian blending of the LDR CMOS image with the ESRGAN image. Again, this method struggles from the same limitations of ESRGAN where high frequency artifacts appear in the over-exposed image regions (clouds in the sky and brightly lit mountain-side.) Our proposed method performs consistently better and produces images that best resemble the ground truth. However, when dealing with a large overexposed area containing high-frequency features, our method tends to output slightly softer images, as indicated by the red color (Suppl. Fig. 4(b)).

Conclusions: Suppl. Figs. 1 and 2 illustrate that our method is a versatile HDR imaging method capable in a wide variety of challenging imaging scenarios, and the output images have extremely high-dynamic range, while being free of artifacts, true to the appearances of ground truth, and visually appealing.



Supplementary Figure 4. **HDR-VDP quality maps:** We show spatial quality maps generated by the HDR-VDP algorithm for various baselines. Red color indicates low scores with large perceptible difference from the ground truth (when the images are displayed on an HDR display) whereas green and blue regions indicate little or no perceptible difference from ground truth.

S. 2. RGB vs. Monochrome Inputs on Baselines



Supplementary Figure 5. **CMOS RGB Inputs vs. Monochrome Inputs:** Visual comparison of 8-bit RGB CMOS input and 32-bit monochromatic (green channel only) CMOS input. The only two baselines that use 8-bit RGB input, DHDR [56] and ExpandNet [44], are highlighted in brown. RGB input provides additional color data, giving the two baselines an advantage. However, our method still significantly outperforms both baselines.

In this section we compare the monochromatic and RGB CMOS inputs. Our proposed method and most of the baselines are given 32-bit monochromatic (green channel only) CMOS input, with pixel values denoting estimated photon flux. However, DHDR [56] and ExpandNet [44] are given 8-bit RGB CMOS inputs.

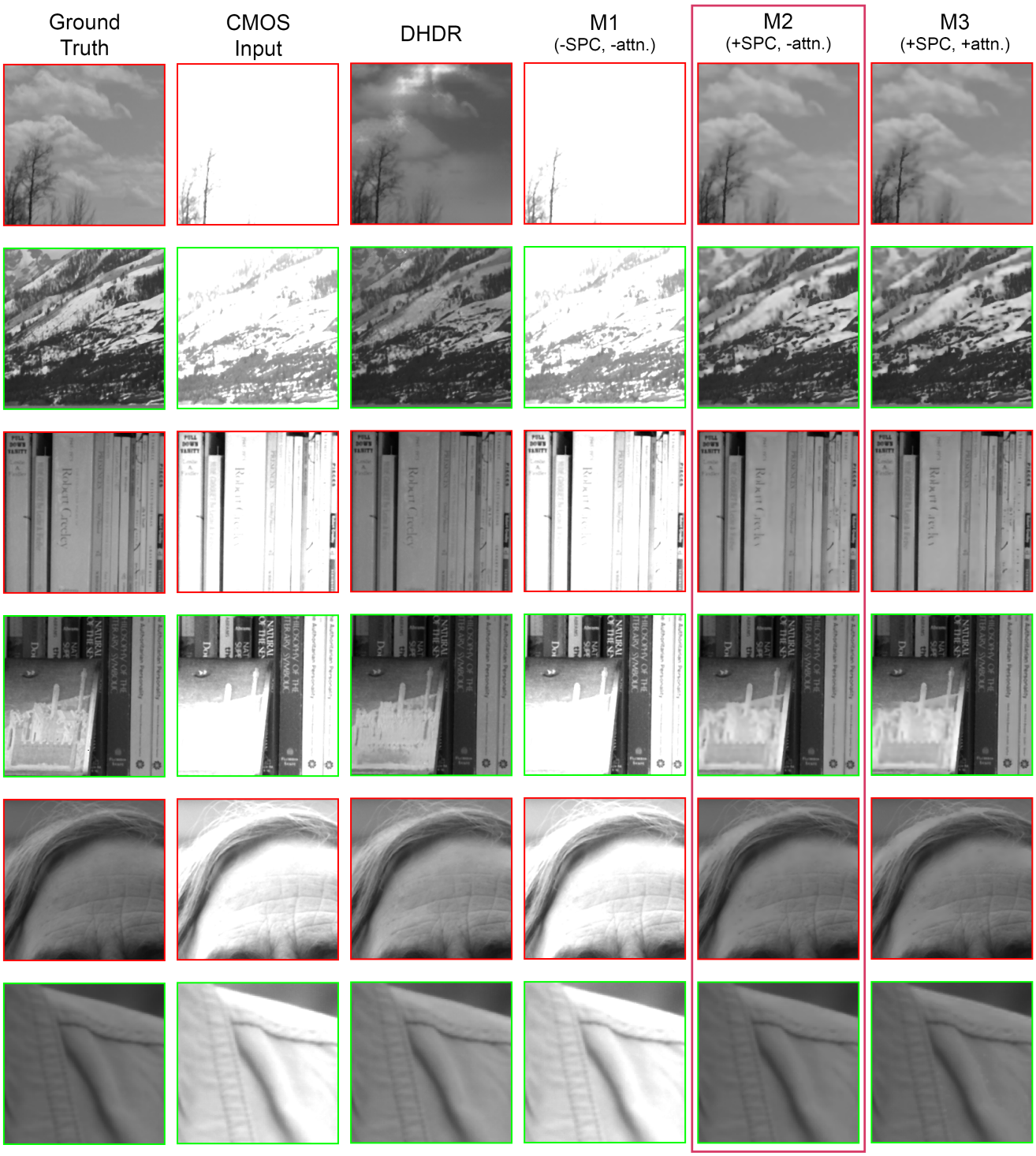
We present the arguments for giving 8-bit RGB CMOS inputs to DHDR and ExpandNet in Suppl. Fig. 5:

- **Additional color information**: as seen in row 1 in Suppl. Fig. 5, the sky region for this scene saturates in green channel, however, blue channel in the RGB CMOS image still provides additional data, such as the shape of the clouds. Such additional data is critical for single-image HDR networks, for instance, DHDR leverages information from non-saturated color channels and contextual information to synthesize textures. Moreover, since our testing dataset is poorly white balanced and most images contain a green color cast (as seen in Suppl. Fig. 5), green channel readily saturates before red and blue, therefore giving the two single-image baselines more unsaturated information to work with. As a counter example, in row 3, none of the three channels provide valid information on the the back of the truck, and as the result, neither single-image baseline is able to recover any texture in this region.
- **Limited sacrifice on dynamic range:** We use linear re-scaling to compress the 32-bit CMOS images into the range of [0, 255]. While such compression inevitably causes some loss of bit precision, however, since all CMOS simulations are already constrained by the 15-bit full well capacity limit, we find no visible differences between the 32-bit .hdr files and the 8-bit PNG files, especially in the highlights[3]. Given that this paper mainly concerns with extending dynamic range towards highlights (as opposed to shadows), such 8-bit compression results in little loss of quality.
- **Replicating the distribution of training data:** Both DHDR and ExpandNet are trained on 8-bit CMOS images. We previously attempted feeding in 32-bit input files (RGB and monochrome) into the two baselines, but these yielded poor results, suggesting that our 32-bit data is outside the acceptable range of inputs for their networks. We find 8-bit inputs, especially 8-bit RGB inputs, yield much better results from these two baselines.

Overall, giving 8-bit RGB inputs to DHDR and ExpandNet is not only necessary but also results in a comparative advantage over monochrome inputs. Nonetheless, as seen in Suppl. Fig. 5, our method still significantly outperforms both baselines.

[3]In Photoshop, visual differences (mainly quantization artifacts) starts to become apparent only when a sufficiently high exposure adjustment filter is applied (greater than, say, 10 stops)

S. 3. Ablation Study



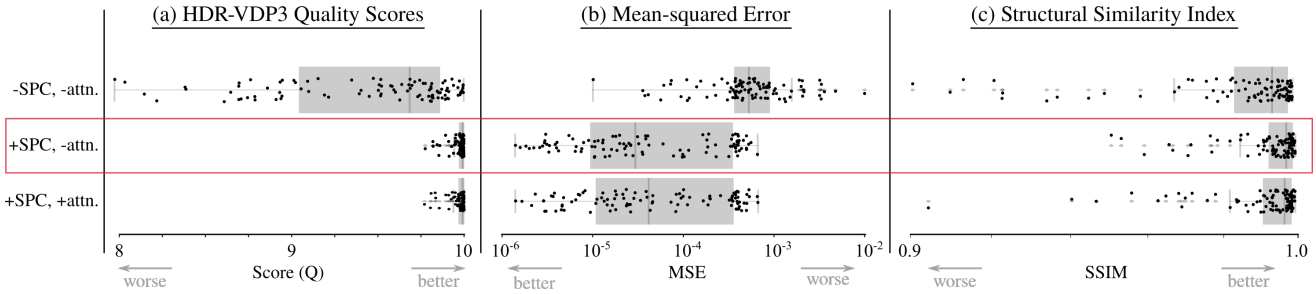
Supplementary Figure 6. **Ablation Study Results:** Visual comparison between our proposed model (M2, highlighted in the red rectangular) and two ablation variants: a model without SPC data (M1), and a model with additional attention masks (M3). The addition of SPC data (M2) yields significant improvements in visual quality over (M1). However, additional attention masks (M3) only result in minimal differences in the shapes and boundaries of recovered features, which is nearly indistinguishable. Therefore, we choose (M2) as our proposed model for the reduced network complexity.

In this section, we present a visual comparison for our ablation models. As described in the main text, we have three candidate models: (M1) is the U-net backbone that takes on the LDR CMOS input, (M2) takes SPC information through an encoder and (M3) introduces additional attention modules in the decoder side of the full network.

Performance of (M1): We investigate the effectiveness of SPC data. We modify our network by removing SPC inputs and the corresponding encoder. Without the guided SPC data, the network essentially acts as a single-image HDR U-net and its performance is generally worse than all single-image HDR baselines, since it is not optimized to work only with CMOS data.

Performance of (M2): Introducing the SPC input and corresponding encoder branch in the network gives significant improvement in visual quality and quantitative metrics, as seen in Suppl. Figs. 6 and 7.

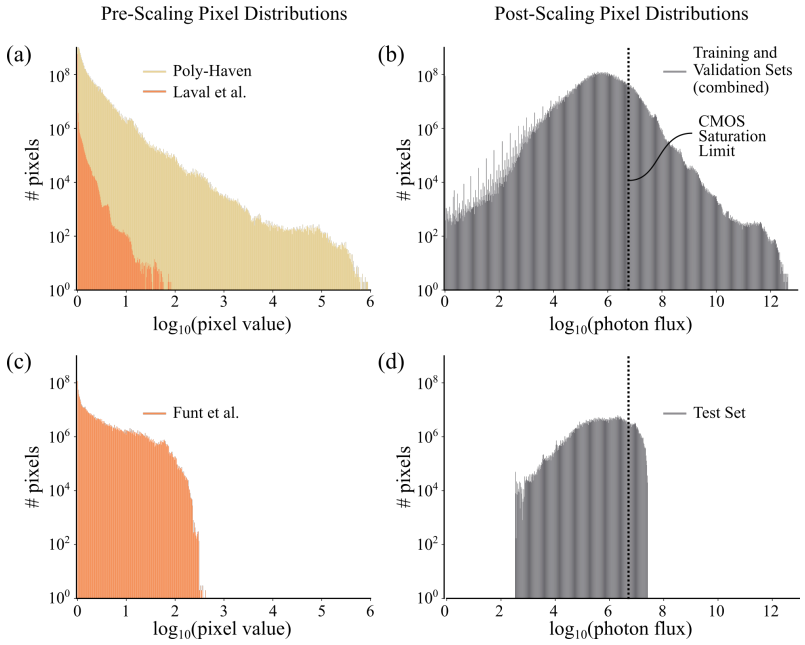
Performance of (M3): Some image neural networks use attention masks to filter out invalid information [25]. We add attention mask modules at two different places, each of which resides at the concatenating layers between the CMOS encoder and the main decoder. These attention masks attempt to filter out over-exposed regions from the CMOS input. In general, without the attention modules, the results have sharper and more defined boundaries on the saturated regions, while the model with attention masks produces smoother edges on the saturated regions. Other than these minor differences, we find that with attention masks, the performance is nearly indistinguishable from our proposed method in terms of visual quality and numerical metrics. Therefore we choose (M2) as the final proposed model.



Supplementary Figure 7. **Numerical metrics for ablated models:** (a) HDR-VDP quality metric, (b) MSE and (c) SSIM computed with μ-compressed images ($\mu = 500$). Introducing SPC information provides the largest performance improvement with all three quality metrics. Introducing the additional attention gates does not improve performance enough to warrant the additional network complexity. Therefore our proposed model (highlighted in the red rectangle) consists only of the U-net backbone and the SPC encoder without any attention modules.

S. 4. Pixel Values of Synthetic Dataset

For training, validation, and testing of our model we gathered three datasets: Poly Haven [28] (train/val), Laval et al. [20] (train/val), and Funt et al. [18, 19] (test). Suppl. Fig. 8 shows the histograms for the unscaled and scaled per-pixel brightness values of all images from each dataset. We apply linear scaling to the three datasets, and we show the histograms of the scaled datasets in Suppl. Fig. 8. We show that our scaled datasets contains a large percentage of pixels that are saturated in the simulated CMOS images [4], indicated by the vertical dashed lines. However, Funt et al. contains substantially fewer scenes with extreme DR (e.g., shooting directly into the sun in a sunny day), the scaled test set is therefore less challenging than our scaled train/dev set, a condition that is more favorable for the single-image baselines (DHDR [56] and ExpandNet [44]) and Dual-Exposure Bracketing. Nonetheless, our synthetic evaluation shows that our proposed method still outperforms these baselines, as seen in Section 6.2 and Supplementary Section S. 1.



Supplementary Figure 8. **Histograms of our synthetic datasets:** Histograms of pixel values for images used in our synthetic evaluation before sclaing (a, c) and after scaling (b, d). The dashed line indicates the CMOS saturation limit on the scale of log photon flux in our scaled datasets.

The datasets are scaled such at after scaling the smallest and largest flux values map to a range of $10^1 - 10^{12}$ photons/second. This is based on photon flux calculations[5] that assume a $50\,\mu\mathrm{m}$ SPC pixel size, $f/D = 2.8$, array fill-factor of 50%, incident light at green wavelength at $555\,\mathrm{nm}$, and luminance values in real-world scenes[6] that span $10^{-6} - 10^6$ cd/m^2.

[4]The flux level that corresponds to the CMOS saturation limit is equal in all synthetic scenes shown in Supple. Fig. 8 because the CMOS parameters (e.g., full well capacity and exposure time) are held constant for all images.

[5]Schöberl et al. "Photometric limits for digital camera systems." Journal of Electronic Imaging 21.2 (2012), Eqs. (1)-(4) and (7).

[6]"Lumens and mesopic vision" Application Note FAQ0016-0297 Osram Sylvania Corp. (2000)