# Supplementary material for Single image HDR reconstruction

Param Hanji
param.hanji@cl.cam.ac.uk
University of Cambridge
United Kingdom

Rafał K. Mantiuk
University of Cambridge
United Kingdom
rafal.mantiuk@cl.cam.ac.uk

Gabriel Eilertsen
Linköping University
Sweden
gabriel.eilertsen@liu.se

Saghi Hajisharif
Linköping University
Sweden
saghi.hajisharif@liu.se

Jonas Unger
Linköping University
Sweden
jonas.unger@liu.se

This is the supplementary document for the paper "Comparison of single image HDR reconstruction methods — the caveats of quality assessment" (DOI: 10.1145/3528233.3530729). Additional materials, including an interactive HTML result viewer, can be found at the project website[1].

Here, we include additional figures and tables. Some figures are extended versions of the main figures showing plots for more scenes or metrics, while values from others are included in the text of the main document.

## 1 SUBJECTIVE IMAGES AND DATA

To run the subjective pairwise comparison experiment, we selected a diverse subset of 27 scenes depicted in Figure 2 in the main document. Here, in Figure 1, we show the results of the experiment for each scene and the two exposures corresponding to 5% and 3% clipped pixels. Notice that both exposures show similar rankings of the methods for each scene, but there is a large variance in subjective quality between scenes.

## 2 ADDITIONAL RESULTS

### 2.1 Per-content image quality predictions

The performance of a quality metric is typically evaluated by checking the correlation with the subjective data. This is typically done per-content and per-method. We report the resulting Spearman rank order coefficients of such a per-content predictions for all tested metrics in Table 1, both before and after polynomial correction. Unfortunately, none of the coefficients are satisfactory as the metrics are not well-suited to judge the absolute per-content quality of the reconstructions of SI-HDR methods. However, they fare better when we average the quality predictions across multiple images, as shown in the main paper.

**Table 1: Correlation coefficients (Spearman) between metric predictions and subjective data both without and with CRF inversion, computed for individual conditions.**

| Metric | Without CRF correction | With polynomial CRF correction |
|---|---|---|
| PU21-PSNR | 0.472 | 0.4 |
| PU21-SSIM | 0.314 | 0.303 |
| PU21-LPIPS | 0.397 | 0.394 |
| HDR-VDP-2 (Q) | 0.237 | 0.434 |
| HDR-VDP-3 (Q) | 0.452 | 0.548 |
| FovVideoVDP | 0.466 | 0.366 |
| PU21-MS-SSIM | 0.425 | 0.482 |
| PU21-FSIM | 0.473 | 0.55 |
| PU21-VSI | 0.438 | 0.451 |
| PU21-VIF-HDR | 0.393 | 0.388 |
| PU21-VIF-SDR | 0.348 | 0.324 |
| PU21-PIQE | 0.13 | 0.0791 |
| PU21-BRISQUE | 0.0771 | 0.0405 |
| PU21-NIQUE | 0.0632 | 0.0158 |
| PU21-NIMA | 0.377 | 0.14 |
| $\mu$-PSNR | 0.372 | 0.373 |
| $\mu$-SSIM | 0.343 | 0.304 |
| $\mu$-LPIPS | 0.4 | 0.353 |
| $\mu$-VSI | 0.396 | 0.402 |
| $\mu$-PIQE | 0.129 | 0.0799 |
| Reinhard-PSNR | 0.393 | 0.357 |
| Reinhard-SSIM | 0.398 | 0.26 |
| Reinhard-LPIPS | 0.332 | 0.334 |
| Reinhard-VSI | 0.391 | 0.406 |
| Reinhard-PIQE | 0.109 | 0.0791 |
| Linear-PSNR | 0.35 | 0.271 |
| Linear-SSIM | 0.487 | 0.375 |
| Linear-LPIPS | 0.183 | 0.16 |
| Linear-FSIM | 0.513 | 0.476 |
| Linear-VSI | 0.256 | 0.277 |
| Linear-PIQE | 0.13 | 0.0791 |

### 2.2 The required number of images

In the main paper we show that averaging quality scores across multiple images improves the accuracy of quality predictions. Here, we show how the number of images affects such predictions.

We generated 2000 bootstrapped samples, where each sample simulated the outcome of an experiment since it considered a different set of observers and different scenes (sampling with replacement). We varied the number of images used for the bootstrapping procedure from 5 to 54 (27 scenes $\times$ 2 clipping levels). From Figure 2, we see that the correlations of most metrics plateau between 20 and 30 images. The proposed polynomial correction (Figure 2b) improves the scores of all but a few no-reference metrics.

### 2.3 Metric comparison – statistical significance

We performed statistical tests to determine which metrics can be said to perform better (in terms of correlation) at 95% confidence level. Because the distributions were non-normal (after Fisher's transform), we performed a non-parametric test by directly computing the distribution of the difference of bootstrap samples and report the $5^{\text{th}}$ percentile in Table 2. The results show that, despite a sensitive experimental protocol with a large number of images and participants, we cannot rank with 95% confidence the top eight metrics (from PU21-VSI to PU21-NIQUE in Figure 5b in the main document). However, we can exclude the metrics that are very

[1]Project web page: https://www.cl.cam.ac.uk/research/rainbow/projects/sihdr_benchmark

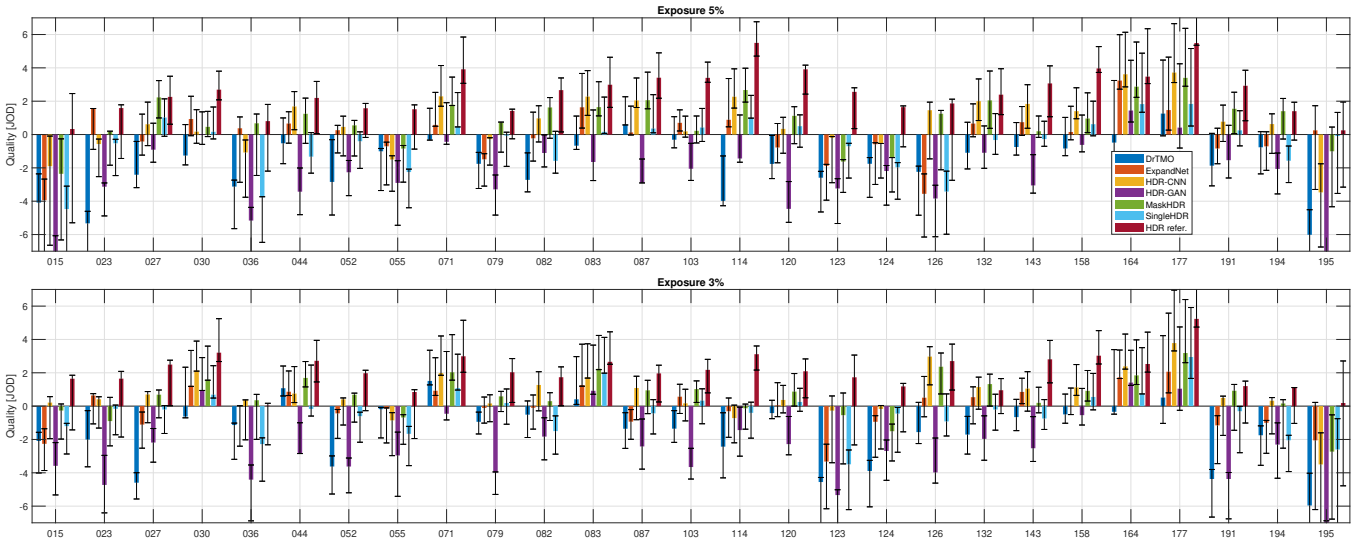
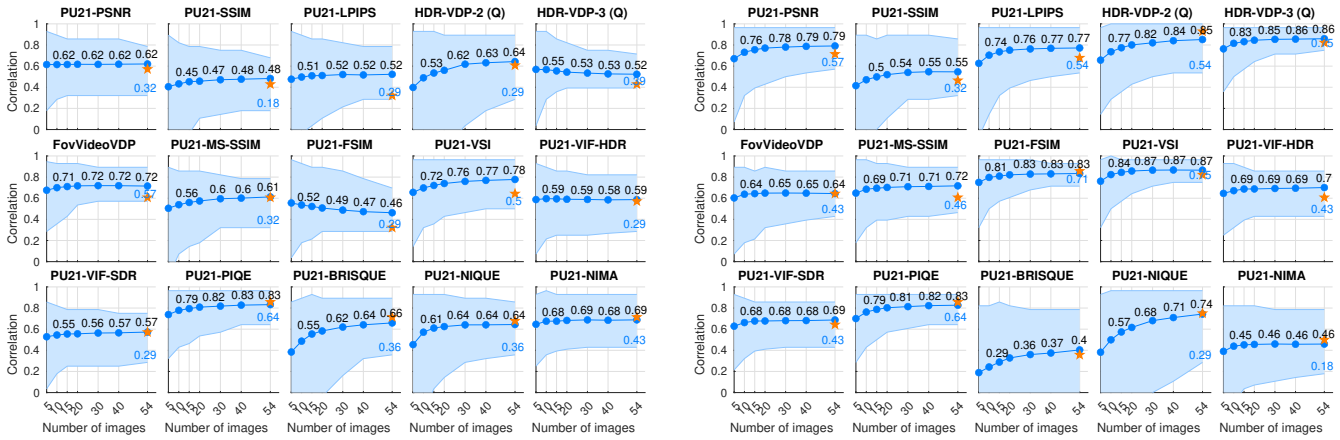



**Figure 1: Preference of the SI-HDR method results shown per exposure (top/bottom) and for all content (x-axis). The error bars show bootstrapped 95% confidence intervals.**




**(a) Without CRF correction** **(b) With polynomial CRF correction**

**Figure 2: The correlation of the averaged quality metrics with the aggregated experiment results, both without (left) and with polynomial CRF correction(right). The results are bootstrapped for different number of images (x-axis). The shaded region represents 95% confidence interval. The orange stars show the results using all the images and all observers (no bootstrapping).**

likely to perform worse, including PU21-SSIM and PU21-NIMA (commonly used in SI-HDR evaluations).

## 2.4 Adaptation of SDR metrics to HDR

SDR metrics cannot be directly used on (linear) HDR pixel values and instead they need to be adapted to the HDR data. Several such adaptations have been used in SI-HDR papers. Here we compare PU21 transform [Mantiuk and Azimi 2021], $\mu$-transform [Kalantari and Ramamoorthi 2017] or Reinhard et al. global tone-mapping operator [Reinhard et al. 2002] and also the incorrect practice of using SDR metric on linear values. From the results in Figure 3 and Figure 4, we did not find evidence for a significant difference

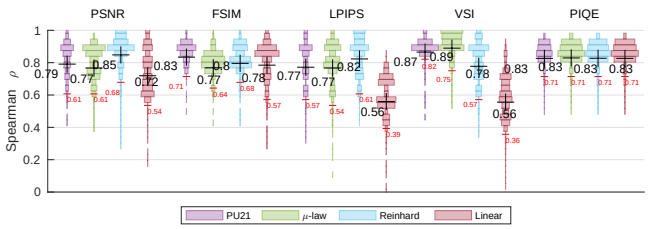



**Figure 3: Comparison of methods used to adapt SDR metrics to HDR images. The incorrect practice of using no-adaptation or tone-mapping can degrade metric performance.**

**Table 2: Statistical test to compare ranked metrics. 1 indicates that difference between the metrics is statistically significant at the 5% significance level.**

| | PU21-VSI | HDR-VDP-3 | HDR-VDP-2 | PU21-FSIM | PU21-PIQE | PU21-PSNR | PU21-LPIPS | PU21-NIQUE | PU21-MS-SSIM | PU21-VIF-HDR | PU21-VIF-SDR | Fov-VideoVDP | PU21-SSIM | PU21-NIMA | PU21-BRISQUE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PU21-VSI | - | - | - | - | - | - | - | - | - | 1 | 1 | 1 | 1 | 1 | 1 |
| HDR-VDP-3 | - | - | - | - | - | - | - | - | - | 1 | 1 | 1 | 1 | 1 | 1 |
| HDR-VDP-2 | - | - | - | - | - | - | - | - | - | - | - | 1 | - | 1 | - |
| PU21-FSIM | - | - | - | - | - | - | - | - | 1 | 1 | 1 | - | 1 | 1 | 1 |
| PU21-PIQE | - | - | - | - | - | - | - | - | - | - | - | 1 | - | 1 | 1 |
| PU21-PSNR | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | 1 |
| PU21-LPIPS | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| PU21-NIQUE | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| PU21-MS-SSIM | - | - | - | 1 | - | - | - | - | - | - | - | - | - | - | - |
| PU21-VIF-HDR | 1 | 1 | - | 1 | - | - | - | - | - | - | - | - | - | - | - |
| PU21-VIF-SDR | 1 | 1 | - | 1 | - | - | - | - | - | - | - | - | - | - | - |
| FovVideoVDP | 1 | 1 | 1 | - | 1 | - | - | - | - | - | - | - | - | 1 | - |
| PU21-SSIM | 1 | 1 | - | 1 | - | 1 | - | - | - | - | - | - | - | - | - |
| PU21-NIMA | 1 | 1 | 1 | 1 | 1 | - | - | - | - | - | - | 1 | - | - | - |
| PU21-BRISQUE | 1 | 1 | - | 1 | 1 | 1 | - | - | - | - | - | - | - | - | - |

**Table 3: Statistical test to compare encoding curves for selected SDR metrics. Rows and columns of each table are ranked according to mean correlation. 1 indicates that difference is statistically significant at the 5% significance level.**

**(a) PSNR**

| | Rein. | PU21 | $\mu$ | Linear |
|---|---|---|---|---|
| Rein. | - | - | - | - |
| PU21 | - | - | - | - |
| $\mu$ | - | - | - | - |
| Linear | - | - | - | - |

**(b) FSIM**

| | PU21 | $\mu$ | Linear | Rein. |
|---|---|---|---|---|
| PU21 | - | - | - | 1 |
| $\mu$ | - | - | - | 1 |
| Linear | - | - | - | 1 |
| Reinh. | 1 | 1 | - | - |

**(c) LPIPS**

| | Rein. | PU21 | $\mu$ | Linear |
|---|---|---|---|---|
| Rein. | - | - | - | 1 |
| PU21 | - | - | - | 1 |
| $\mu$ | - | - | - | 1 |
| Linear | 1 | 1 | 1 | - |

**(d) VSI**

| | PU21 | $\mu$ | Rein. | Linear |
|---|---|---|---|---|
| PU21 | - | - | - | 1 |
| $\mu$ | - | - | - | 1 |
| Rein. | - | - | - | 1 |
| Linear | 1 | 1 | - | - |

**(e) PIQE**

| | $\mu$ | Rein. | PU21 | Linear |
|---|---|---|---|---|
| $\mu$ | - | - | - | - |
| Rein. | - | - | - | - |
| PU21 | - | - | - | - |
| Linear | - | - | - | - |

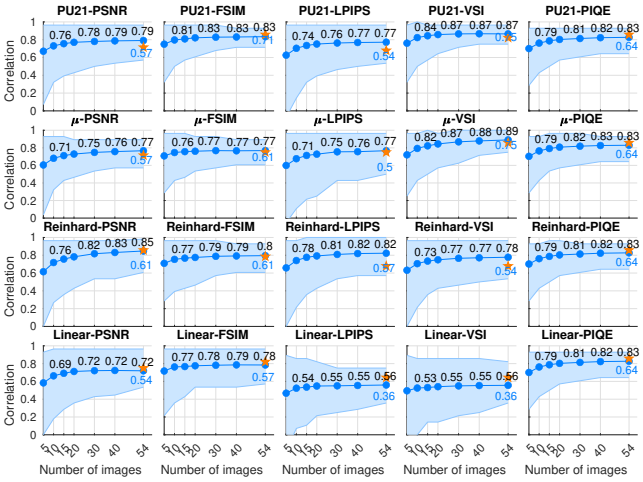



**Figure 4: Comparing the different HDR-to-standard dynamic range (SDR) transforms PU-21, $\mu$-law, global tone mapping against the base case which is an identity mapping resulting in linear pixel values. The plots show bootstrapped correlations with subjective scores over the experiment subset.**

between PU21 and $\mu$-law. But we noted that computing metrics on linear values almost always results in significantly worse performance. Using tone mapping also reduced performance of several metrics. This is confirmed by the statistical tests in Table 3. Average correlations with 95% confidence intervals for all metrics are shown in Figure 4.

## 2.5 Validation on test dataset

We validate our proposed protocol on the rest of the SI-HDR data consisting of 156 images at exposures 3% and 5%. Figure 5 depicts scatter plots of JODs vs metric predictions for the test data. We also show bootstrapped results to the rank the metrics. This is the same as Figure 6 from the main document but contains more metrics.

## 2.6 Confidence intervals for image quality distributions

If image quality scores are averaged across many images, we could be tempted to compute confidence intervals on the distribution of quality scores across the images. Such confidence intervals, however, do not ensure that the differences are large enough to claim better performance. To show an example, we bootstrapped the selection of images and compute the metrics for 2000 bootstrap samples, from which we extract 95% confidence intervals, shown in Figure 6 in the main paper. While the confidence intervals are not a substitute for a statistical test, the large differences clearly indicate statistically significant differences between many conditions, which were incorrectly ranked. This is because bootstrapping the samples of images accounts only for the measurement error due to the selection of images. It does not account for the measurement error due to the inaccuracy of the metric and the inaccuracy of the subjective data.

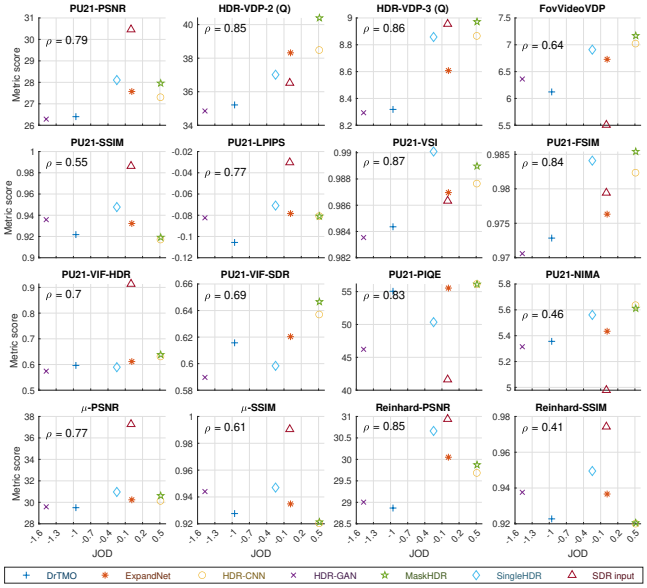



**Figure 5: Validation of the metric predictions on the unseen portion of the dataset. The scatter plots show the mapping from JOD, measured for 54 conditions in the experiment, to the average metric predictions, computed on the remaining portion of the dataset. $\rho$ is the mean Spearman's rank correlation coefficient computed using the unbiased estimator [Olkin and Pratt 1958]. In all plots, input JOD has been scaled to 0. Different colors depict different SI-HDR methods.**

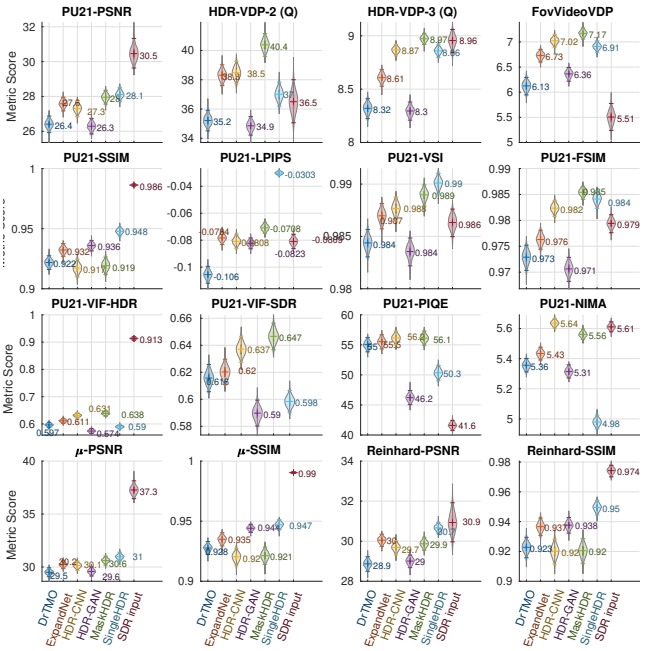



**Figure 6: Rankings of all methods on 2000 bootstrapped sets drawn from 156 images not used in the experiment. Different colors depict different reconstruction methods and error bars show 95% confidence interval.**

# REFERENCES

Nima Khademi Kalantari and Ravi Ramamoorthi. 2017. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.* 36, 4, Article 144 (jul 2017), 12 pages. https://doi.org/10.1145/3072959.3073609

Rafał K. Mantiuk and Maryam Azimi. 2021. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *Picture Coding Symposium*. 1–5.

Ingram Olkin and John W. Pratt. 1958. Unbiased Estimation of Certain Correlation Coefficients. *The Annals of Mathematical Statistics* 29, 1 (1958), 201 – 211. https://doi.org/10.1214/aoms/1177706717

Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. 2002. Photographic Tone Reproduction for Digital Images. *ACM Trans. Graph.* 21, 3 (jul 2002), 267–276. https://doi.org/10.1145/566654.566575