Supplementary for: "Consolidated Dataset and Metrics for High-Dynamic-Range Image Quality"

Aliaksei Mikhailiuk, María Pérez-Ortiz, Dingcheng Yue, Wilson Suen, and Rafał K. Mantiuk

I. Introduction

THIS supplementary file contains additional information that we could not include in the main paper due to space considerations. Below we describe: (i) the experimental procedure for the cross-dataset and within dataset pairwise comparisons; (ii) justification for the linear complexity of the model for psychometric scaling; (iii) detailed architecture of the PU-PieApp; (iv) selection procedure for datasets included in UPIQ (v) detailed architecture of the multitask network; (vi) Maximum Differentiation (MAD) competition for the tested metrics; (vii) example images from the UPIQ dataset.

II. Experimental Protocol for Dataset Merging

Here we include additional details about the design of the experiments for collecting required cross-dataset and within dataset quality measurements. In order to produce a meaningful unified quality scale using pairwise comparisons for a specific single IQA dataset one needs a) comparisons of distorted to pristine quality reference image, b) within-content comparisons to scale different levels of distortion for the same distortion type and c) cross-content comparisons [1], to connect all content and put them on the same quality scale. For rating this would be equivalent to having observers rate images across all distortions and distortion levels during the same session, instead of having separate experiments. In the case of selected datasets, all of these considerations were taken into account when original data was collected, i.e. each dataset has a self-contained unified quality scale. To align these datasets we need to connect disjoint scales through pairwise comparisons and also find the relationship between rating and pairwise comparison judgments within each of the datasets. This means that for every disjoint rating dataset we need to collect within dataset comparisons and link all datasets with across dataset comparisons.

A. Displays and stimuli

For the presentation on the HDR display we transform all images from either gamma-corrected (SDR) or relative linear (HDR) pixel value to absolute linear colorimetric units in the Rec. 709 colour space [2]. The peak luminance of the images was matched to the peak luminance of the displays used to collect original datasets. The images were also displayed with the same angular resolution (in pixels per visual degree) as in the original experiments. When the image size exceeded the size of our display, we provided a simple panning interface in which observers could use a trackball to inspect different portion of the image.

A. Mikhailiuk, D. Yue, W. Suen and R. Mantiuk are with the Department of Computer Science and Technology at the University of Cambridge (UK) (email: {am2442, dy276, wss28, rkm38}@cam.ac.uk).

M. Pérez-Ortiz is with the Department of Computer Science at the University College London (UK) (email: maria.perez@ucl.ac.uk)

B. Experimental Procedure and Participants

We extended the data collected in original datasets and follow-up studies for TID2013 [3], [4] and LIVE [5] datasets with two additional pairwise comparison experiments. In all cases, comparisons to be performed were selected so that compared images were of similar quality, excluding obvious comparisons so as to maximise informativeness of the collected data. Note that this is a common approach in pairwise comparison experiments and the basis for active sampling approaches [5].

In the first experiment we collected only comparisons within the dataset, *i.e.* comparing images of the same dataset. This is necessary for finding the relationship between rating measurements and pairwise comparisons. It is only necessary for rating-based datasets, which means we excluded TID2013 from this experiment since we used previously collected pairwise comparisons and rating measurements [3], [4]. We ensured that all three types of previously mentioned comparisons were covered: to reference, within-content and cross-content. After the first experiment, all the data could be scaled, since we had comparisons to a common reference for all datasets.

For the second experiment we compared conditions exclusively from different datasets, connecting each dataset to the rest. Images were chosen to uniformly cover the quality scale. We performed several iterations of the pair selection. After conducting a pairwise comparison experiment on a small batch of comparisons, we re-scaled the dataset with newly collected comparisons and selected the next batch from the new scale.

Observers were asked to compare two distorted images and choose the one with better quality with respect to their reference. The reference image could be viewed by pressing and holding a space bar. The observer was asked to see the reference images at least once for each comparison. The order of comparisons in every experiment was randomized. We ensure that ITU recommendations [6] were met. And that the time for performing one experiment did not exceed 30 minutes, so as to prevent observer tiredness from influencing the experiment outcomes. Each selected pair of images was compared by 6 participants, with each participant completing approximately 300 trials. Overall 6000 new comparisons were collected from 20 participants.

III. Relationship between pairwise comparisons and mean opinion scores

Watson [7] studied the correlation between rating scales and results of pairwise comparisons in the context of psychometric scaling of pairwise preference probabilities. He found that the degree of agreement between two scales, for the case of video compression, is relatively high. The work reports a quadratic relationship between MOS and scaled PWC, with a very small quadratic coefficient. On the contrary [1] shows that there is a strong linear relationship between MOS and PWC scaling results. Here we test both assumptions to validate, if the linear relationship is indeed sufficient.

To compare goodness of fit we report adjusted R^2 statistic – R^2_{adj}, which, unlike simple R^2 accounts for the number of model parameters in explaining the variance in the data [8]:

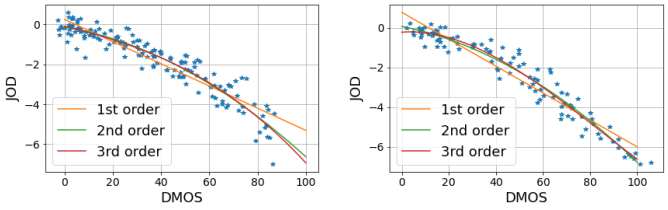
$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{n-p-1}, \tag{1}$$

where R^2 is defined as:

$$R^2 = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y}_i)^2}, \tag{2}$$

where n is the number of data points in the dataset, p is the number of parameters, excluding the constant term, $\mathbf{y}$ and $\hat{\mathbf{y}}$ true and predicted response variables and $\bar{y}$ is the mean of $\mathbf{y}$.

We fit 1^{st} 2^{nd} and 3^{rd} order polynomials into the JOD, obtained from pairwise comparisons, and MOS of TID2013 [9] and LIVE [10] image quality datasets. Figure 12 shows the scatter plot of the scores and fitted polynomials. An important observation can be drawn – the model describing the relationship between JOD and MOS for image quality must be monotonic, as an increase/decrease in the quality of an image should result in the increase/decrease of the scores in both scales. Violation of this requirement is visible in the example of 3^{rd} order fit into the scores from LIVE dataset.



(a) TID2013 (b) LIVE

Fig. 12: Polynomial fits into the JOD and MOS scores of three subjective image and video quality datasets.

The results of computing R^2_{adj} are given in Table I. There is only slight increase in R^2_{adj} for TID2013 and LIVE datasets for 2^{nd} and 3^{rd} order polynomials. Non-linear relationship is thus hard to justify given the need for additional constraints on the function to be monotonic.

IV. UPIQ dataset selection

To ensure the accuracy of the data in the UPIQ dataset, candidate datasets were screened with a pilot experiment. We ran a series of within-dataset pairwise comparisons to verify if the ranking of the scores elicited from our subjective study is consistent with the ranking provided in the dataset. Where the scores in the original dataset have shown to have little or no correlation with the data in our experiment, the dataset was not included in UPIQ. Figure 13 shows the scatter plot of the empirical probabilities found in our subjective experiment for a set of image pairs, compared between six and ten times, versus their difference in the MOS scale obtained by [11].

We verify if the data collected by [11] or us is more consistent we run an additional Maximum Differentiation (MAD) experiment [12]. The pairs of images with the most inconsistent scores are given in Figure 14. One of the advantages of the experimental procedure employed for the data collection in UPIQ is the ability to flip between reference and test image during the experiment. Observers were thus, particularly sensitive to the JPEG blocking artefacts in the large smooth areas of the skies of the lake and sunset images.

TABLE I: R^2_{adj} statistic for polynomial fits describing the relationship between MOS and JOD.

Dataset	1^{st} order	2^{nd} order	3^{rd} order
TID2013	0.77	0.79	0.79
LIVE	0.87	0.89	.89

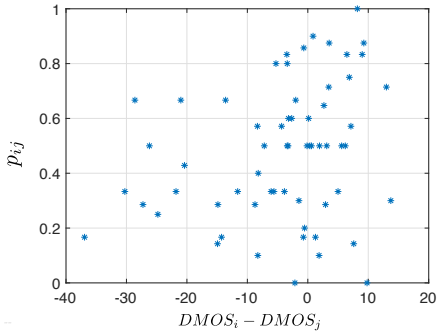
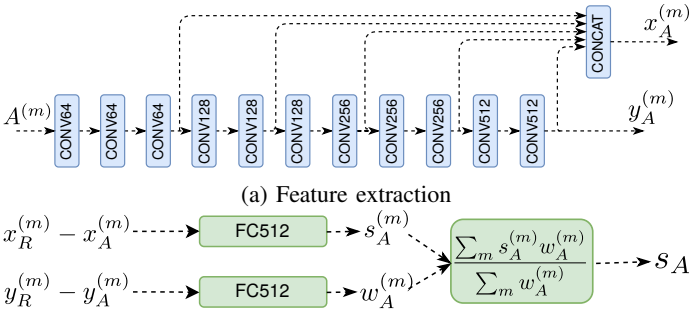


Fig. 13: Scatter plot of the empirical probability p_{ij} obtained from our experiment and difference in the DMOS scores obtained by [11]. Two scales have little correlation with the SROCC of 0.27.

V. PieAPP Detailed Architecture

For every input patch m of reference R and distorted A images the feature extraction (FE) network has two outputs: $y^{(m)}$ from the input passing through the whole network and $x^{(m)}$ formed by concatenation of the flattened outputs of layers at different depths of the network. The score computation (SC) network takes two inputs: the difference between $x_R^{(m)} - x_A^{(m)}$, which is passed through a fully connected layer, predicting patch-wise error s^m and the difference $y_R^{(m)} - y_A^{(m)}$, which is passed through another fully connected layer, producing the patch-wise weight $w^{(m)}$. The two outputs $s^{(m)}$ and $w^{(m)}$, are then used to produce the weighted average of all per patch scores – a quality score of the entire image s_A. Note, that passing two reference images through the network will result in the $x_R^{(m)} - x_A^{(m)} = 0$, thus the output of the

Fig. 14: Representation of image pairs where collected pair-wise comparisons and original DMOS scores from [11] disagree. In each pair the image on the left is the one which has higher DMOS and lower quality from our experiment and the other way around on the right.



(a) Feature extraction

(b) Score computation

Fig. 15: (a) The feature extraction network takes image patches as an input and has two outputs: one from a patch passing through the whole network and another formed from skip connection. The network has 11 convolutional layers with 2×2 max-pooling after every even layer. (b) The score computation network computes patch-wise weights and scores, the weighted average produces the final score

quality estimation function $f(A, B)$, will be constant, defined by the bias of the score computation network. The detailed architecture of the PieAPP network is shown in Figure 15.

a) Alternative Architectures: We experiment with a number of CNN architectures to find the one that generalizes the best. Since the CNN-based metric can be trained end-to-end, it could potentially learn the PU-transform. We replaced the PU-transform with a logarithmic function followed by scaling to the 0-1 range and then trained the network. The prediction error was much higher for the logarithmic function (root-mean-squared error (RMSE) 0.68) compared to the PU-transform (RMSE 0.47). This confirms that the PU is beneficial for quality predictions in SDR/HDR images even for CNN-based metrics.

TID2013 LIVE	Narw. Korsh.	TID2013 Narw.	LIVE Narw.	TID2013 Korsh.	LIVE Korsh.
0.46	0.27	0.33	0.46	0.26	0.10

TABLE II: SROCC between the difference in quality scores $s_A - s_B$, where A and B are images from different datasets and empirical probability p_{ij} for the multitask network.

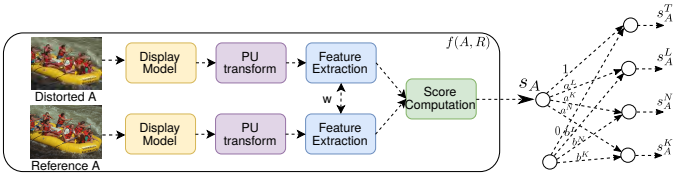


Fig. 16: Multitask network. The network is trained to predict original scores from individual datasets. Similar to the scaling procedure the network learns the implicit quality s_A and parameters a and b for each dataset. To constraint the scores we set parameters of TID dataset $a^T = 1$, $b^T = 0$.

VI. Multitask Network

Collecting data is time consuming and expensive, hence a method capable of learning the implicit unified quality without the need for additional data is desirable. To verify if our network is capable of learning this implicit quality and cross-dataset relationship, we train the network using a multitask learning approach, where it is assumed that all datasets share the same feature representation for quality but since scales are relative the quality scores might be scaled differently. The architecture of the network is given in Figure 16. The $f(A, B)$ part of the network is the same as PU-PieApp and produces a score s_A, which is assumed to be a unified quality for disjoint datasets. Similar to our scaling procedure from Section 3 of the main paper the scores from individual datasets are linked with the unified s_A via a linear relationship. For example the quality score s_A^L for LIVE dataset would be predicted with $a^L * s_A + b^L$, where a^L and b^L are learnt parameters. These parameters from individual datasets are treated and learnt as individual tasks. Since quality scores are relative, we constraint them by setting parameters of TID2013 dataset $a^T = 1$ and $b^T = 0$. To allow for faster convergence we standardized scores from the separate datasets. The training procedure for the multitask network was the same as for the DPIQM.

Similar to Section 4.3 in the main paper we compute the correlation between the difference in quality scores $s_A - s_B$, where A and B are images from different datasets and empirical probability p_{ij}. The detailed results are given in Table II. Neither of the cross-dataset relationships is well captured by the multitask network.

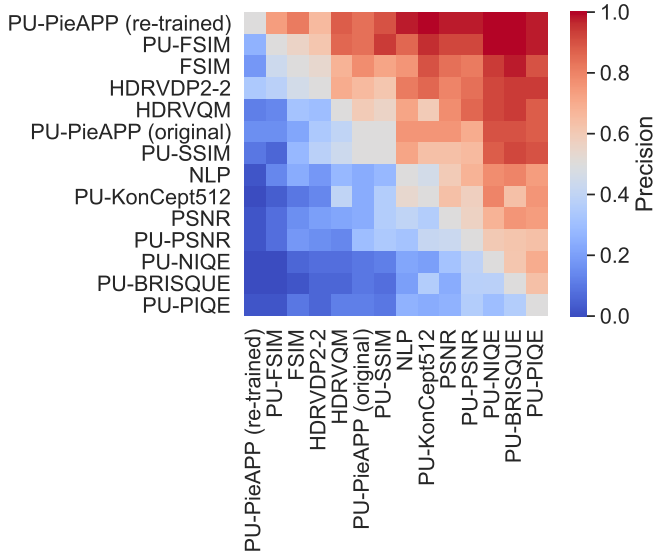


Fig. 17: MAD competition for tested metrics. High values in the row indicate high success of the attack by the metric in the corresponding row on the metric in the corresponding column.

VII. Maximum differentiation competition

We have also performed MAD analysis on the test split of the UPIQ dataset. For a pair of metrics we select pairs of conditions that have different qualities according to the tested quality metric and similar according to the benchmark quality metric. Thus for two quality metrics M^t and M^b with scores in JOD units we select conditions o_i and o_j following $\arg\max_{ij}(|M^t_i - M^t_j| - |M^b_i - M^b_j|)\ subject\ to\ |M^b_i - M^b_j| < 1\ JOD$. Instead of aggressiveness and resistance used in [13], we quantify the performance of a metric by measuring its ability to classify a pair of images as of the same or of different quality. If the absolute difference in JOD units between two images in the UPIQ dataset is < 1, we assume that the conditions are similar in quality, otherwise they are different in quality. We then report precision - the number of pairs correctly ranked and identified as different by the tested quality metric, divided by the total number of selected pairs (100 in our case). The results are given in Figure 17. Each entry of the matrix is the precision of the test metric from the corresponding row when paired against the benchmark quality metric from the corresponding column.

PU-Pie-APP (re-trained) exhibits the best performance both in identifying different in quality (first horizontal row) and similar in quality (first vertical column) conditions when paired with any of the metrics. However, when paired with HDRVDP2-2, PU-PieAPP performs almost on par. Second best performance is attained by PU-FSIM, which has very similar performance to FSIM without PU-transform. Nevertheless, PU-FSIM exhibits stronger performance when paired with metrics accounting for the dynamic range.

VIII. Examples of the dataset

Figures 18 and 19 show sample images from the unified dataset at $JOD = -1$ and -2 and are intended to be a visual subjective validation of the final scale. These levels were selected to show images from all four datasets, as images from the HDR datasets (Korshunov and Narwaria) have quality scores above -2 JOD only. Each figure contains four separated sections, each associated to a different dataset. Each section has two rows: distorted and reference images. For display purposes HDR images were converted to SDR with gamma correction:

$$I_{\mathrm{HDR}} = 255 \left(\frac{I_{\mathrm{SDR}}}{255} \right)^{\frac{1}{2.2}}. \quad (3)$$

As the perceived image quality depends on the display luminance, the SDR images in the figures might be masking or amplifying some image distortions. Thus figures are intended to be an approximate demonstration of the final image quality scale. Nevertheless, images from different datasets at the same JOD level have similar distortion severity. Without our unified photometric image quality dataset (UPIQ) it would be impossible to compare image scores across datasets. Most of the HDR images are distorted only locally, with the overall image quality not deteriorating significantly, as opposed to images from SDR datasets that had uniform distortions applied to them. Narwaria mostly has panorama images, where local distortions are less noticeable due to the size of the image.

References

[1] E. Zerman, V. Hulusic, G. Valenzise, R. K. Mantiuk, and F. Dufaux, "The relation between MOS and pairwise comparisons and the importance of cross-content comparisons," *Proc. of Human Vision and Electronic Imaging*, 2018.

[2] ITU-R, "Parameter values for the hdtv standards for production and international programme exchange," ITU-R Recommendation BT.500-13, Jun 2015.

[3] A. Mikhailiuk, M. Pérez-Ortiz, and R. K. Mantiuk, "Psychometric scaling of TID2013 dataset," in *International Conference on Quiality of Multimedia Experience (QoMEX)*, 2018.

[4] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Transactions on Image Processing*, vol. 29, pp. 1139–1151, 2019.

[5] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4249–4256.

[6] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT.500-13, Jan 2012.

[7] A. B. Watson and L. Kreslake, "Measurement of visual impairment scales for digital video," *SPIE Electronic Imaging, Human Vision and Electronic Imaging VI*, vol. 4299, pp. 79–89, 2001.

[8] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, 1st ed. MIT Press, 2012.

[9] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, and Benoit, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57 – 77, 2015.

[10] H. Sheikh, M. Sabir, and A. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[11] E. Zerman, G. Valenzise, and F. Dufaux, "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience*, vol. 2, no. 1, p. 5, Apr 2017.

[12] W. Zhou and E. P. Simoncelli, "Maximum differentiation (mad) competition: a methodology for comparing computational models of perceptual quantities." in *Journal of Vision*, vol. 8, no. 8.1-13., 2008, pp. 586–595.

[13] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 851–864, 2020.

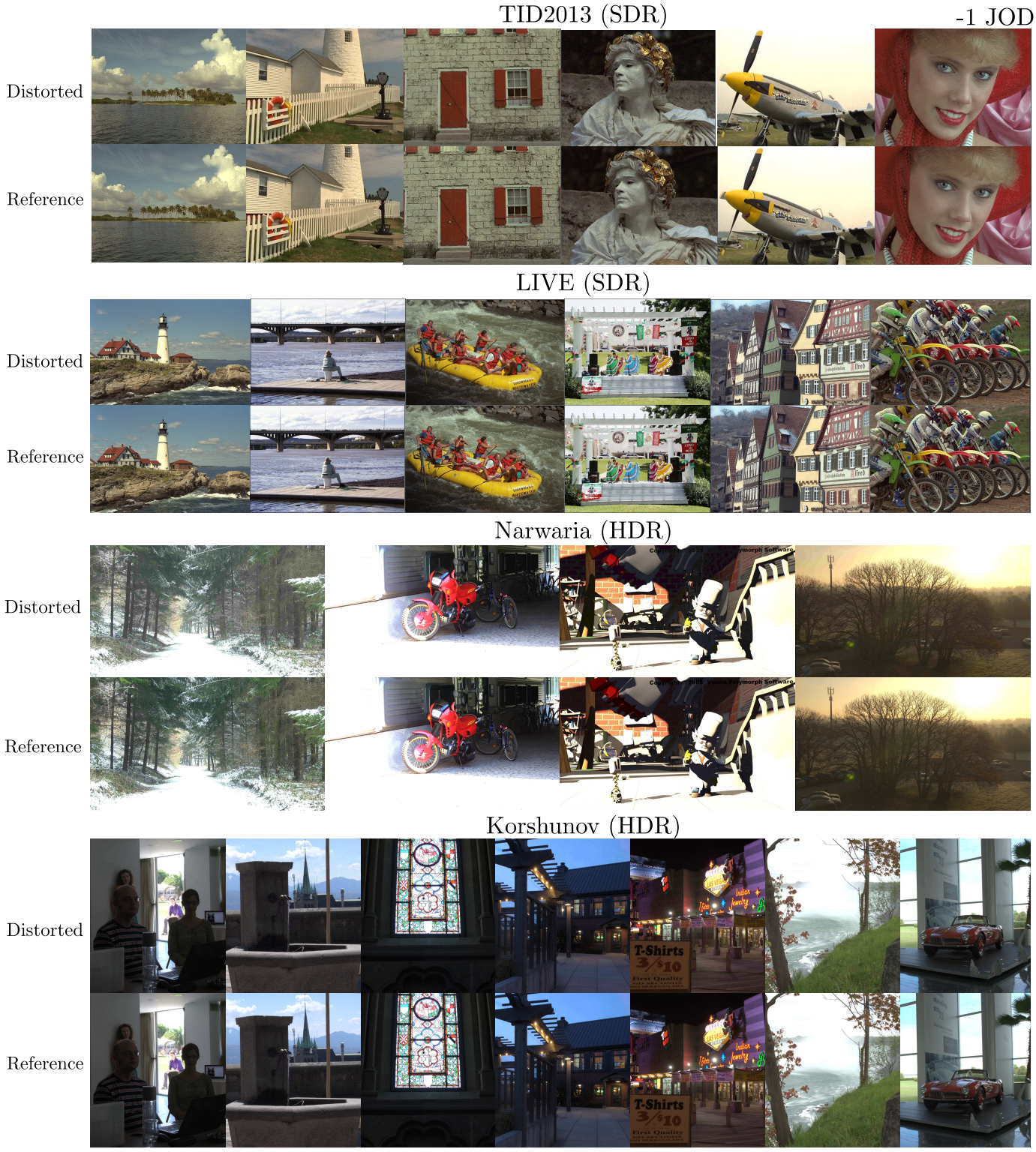


Fig. 18: A selection of images from the four combined datasets at approximately -1 JOD level. Each dataset has two rows: distorted and reference images. We converted HDR images to SDR with gamma correction and gamma 2.2. Images from different datasets at the same JOD level have similar distortion severity. Without a unified dataset it would be impossible to compare image scores across datasets.

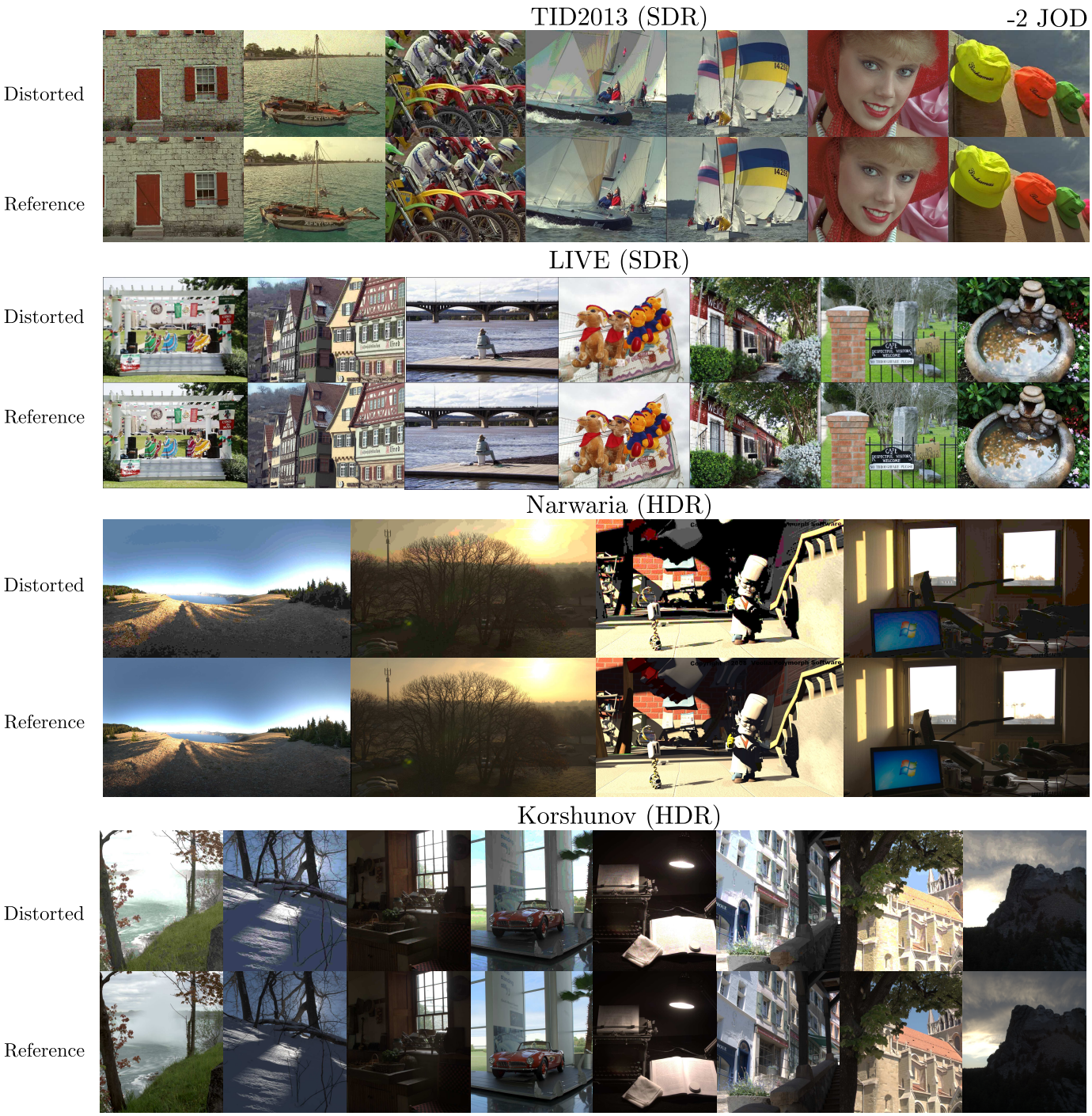


Fig. 19: A selection of images from the four combined datasets at approximately -2 JOD level. Each dataset has two rows: distorted and reference images. We converted HDR images to SDR with gamma correction and gamma 2.2. Images from different datasets at the same JOD level have similar distortion severity. Without a unified dataset it would be impossible to compare image scores across datasets.