FovVideoVDP: A visible difference predictor for wide field-of-view video — Supplementary material

RAFAŁ K. MANTIUK, University of Cambridge
GYORGY DENES, Facebook Reality Labs, University of Cambridge
ALEXANDRE CHAPIRO, Facebook Reality Labs
ANTON KAPLANYAN, Facebook Reality Labs
GIZEM RUFO, Facebook Reality Labs
ROMAIN BACHY, Facebook Reality Labs
TRISHA LIAN, Facebook Reality Labs
ANJUL PATNEY, Facebook Reality Labs

ACM Reference Format:
Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. FovVideoVDP: A visible difference predictor for wide field-of-view video — Supplementary material. *ACM Trans. Graph.* 40, 4, Article 1 (August 2021), 5 pages. https://doi.org/10.1145/3450626.3459831

1 OVERVIEW

This supplementary document contains additional results and analysis for the FovVideoVDP paper [Mantiuk et al. 2021]. This document contains:

- Section 2: Additional plots of the contrast sensitivity function used in the metric.
- Section 3: The description of the masking models compared in the main paper (Figure 16 in the main paper).
- Section 4: Timings of FovVideoVDP compared to other quality metrics.
- Section 6: Additional results and analysis for the FovDots dataset.
- Section 7: Description and results of the merging experiment used to bring multiple datasets into the same quality units.

2 CONTRAST SENSITIVITY FUNCTION

The contrast sensitivity function is typically plotted as a function of frequency. This, however, often obfuscates the fact that other dimensions, such as size and luminance, also have substantial impact on sensitivity. In Figure 1 we plot the CSF as a function of these two dimensions.

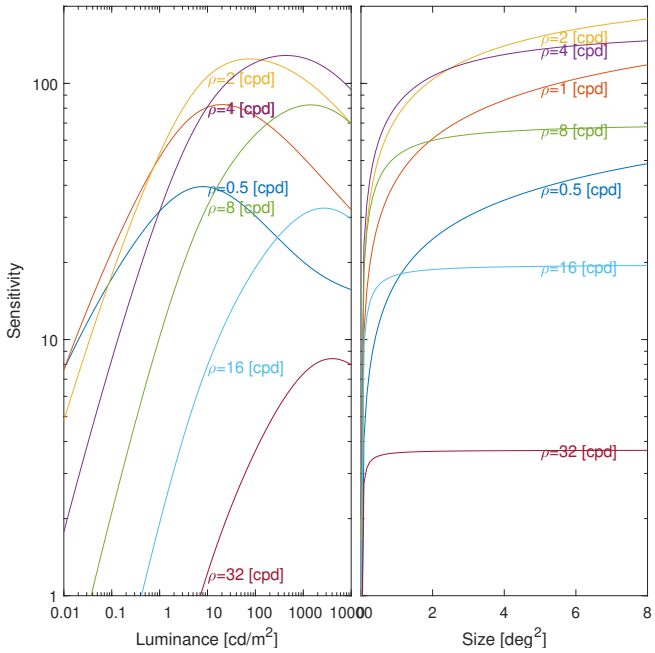

Fig. 1. Luminance contrast sensitivity from [Wuerger et al. 2020] plotted as a function of luminance and size of the stimulus in [deg^2]. The size was set to 3 deg^2 for the left plot and luminance to 100 cd/m^2 for the right plot.

3 CONTRAST MASKING MODELS

We provide additional details on the masking models we used in our ablation study here.

One of the first models that was proposed to explain contrast masking relies on a *contrast transducer* [Foley 1994; Legge and Foley 1980; Watson and Solomon 1997]: a function that transforms physical (Michelson) contrast into the perceived contrast response.

Authors' addresses: Rafał K. Mantiuk, Department of Computer Science and Technology, University of Cambridge, rafal.mantiuk@cl.cam.ac.uk; Gyorgy Denes, Facebook Reality Labs, Department of Computer Science and Technology, University of Cambridge, gyorgy.denes@cl.cam.ac.uk; Alexandre Chapiro, Facebook Reality Labs, alex@chapiro.net; Anton Kaplanyan, Facebook Reality Labs, kaplanyan@gmail.com; Gizem Rufo, Facebook Reality Labs, gizemrufo@gmail.com; Romain Bachy, Facebook Reality Labs, rbachy@fb.com; Trisha Lian, Facebook Reality Labs, tlian@fb.com; Anjul Patney, Facebook Reality Labs, anjul.patney@gmail.com.

0730-0301/2021/8-ART1 $15.00
https://doi.org/10.1145/3450626.3459831

The response of the visual system to a contrast difference is expressed as the absolute difference of the response of two transducers:

$$D_{b,c}(\boldsymbol{x}) = \left| \mathrm{t}(C'^{\,\mathrm{test}}_{b,c}(\boldsymbol{x})) - \mathrm{t}(C'^{\,\mathrm{ref}}_{b,c}(\boldsymbol{x})) \right| , \quad (1)$$

where $C'^{\,\mathrm{test}}_{b,c}$ is the contrast of the test image and $C'^{\,\mathrm{ref}}_{b,c}$ is the contrast in the reference image, both normalized by the sensitivity (Eq.(13) in the main paper). The transducer function models the divisive gain control as:

$$\mathrm{t}(C, S) = \mathrm{sgn}(C) \frac{|C|^p}{Z_0 + (k\,|C| * H)^q} , \quad (2)$$

where the nominator represents the excitatory response of the neurons and the denominator represents the inhibitory signal; p, q and k are the parameters of the model (typically $p = 2.3$, $q = 2$ and $k = 1$ [Watson and Solomon 1997]) and $\mathrm{sgn}(\cdot)$ is the sign function. Z_0 controls the influence of the sensitivity function and is in the range 1-4. H_σ is a convolution kernel (we use a Gaussian) that pools inhibitory signals from neighboring spatial locations. Equation 2 is one of the masking model variants we tested in our ablation study.

Another variant was an adaptation of the threshold elevation function proposed in the book chapter on the VDP metric [Daly 1993]:

$$D_{b,c}(\boldsymbol{x}) = \left(\frac{\left| C'^{\,\mathrm{test}}_{b,c}(\boldsymbol{x}) - C'^{\,\mathrm{ref}}_{b,c}(\boldsymbol{x}) \right|}{T_e(C^{mask}_{b,c}(\boldsymbol{x}))} \right)^p , \quad (3)$$

where the threshold elevation function is defined as:

$$T_e(C^{mask}_{b,c}(\boldsymbol{x})) = \left(1 + \left(k_1 (k_2\, C^{mask}_{b,c}(\boldsymbol{x}))^s \right)^b \right)^{1/b} , \quad (4)$$

the constants are defined as

$$k_1 = W^{1-(1-Q)^{-1}} \quad k_2 = W^{(1-Q)^{-1}-1} , \quad (5)$$

where $W = 6$, $Q = 0.7$ and $b = 4.0$. The slope of the threshold elevation function s was a free parameter that we optimized for the best model fit.

4 TIMINGS

The execution times for Matlab's implementation of FovVideoVDP (both CPU and GPU), and several other metrics are shown in Figure 2. The timings are shown separately for images and video. The execution times were collected on a computer outiftted with an Intel Core i7-7800X CPU and a NVIDIA GeForce RTX 2080 GPU. The reported times are averaged over 5 runs and the standard deviations are shown as error bars.

The GPU implementation of FovVideoVDP is two orders of magnitude faster than HDR-VDP-3, which is another metric that models early-visual perception. It is also faster than most metrics of moderate complexity, although this is mostly because the computations are performed on a GPU while other metrics are computed on the CPU. It should be noted that not every metric can be ported as efficient GPU code. Short execution times are important to enable assessing the quality of high resolution images and video.

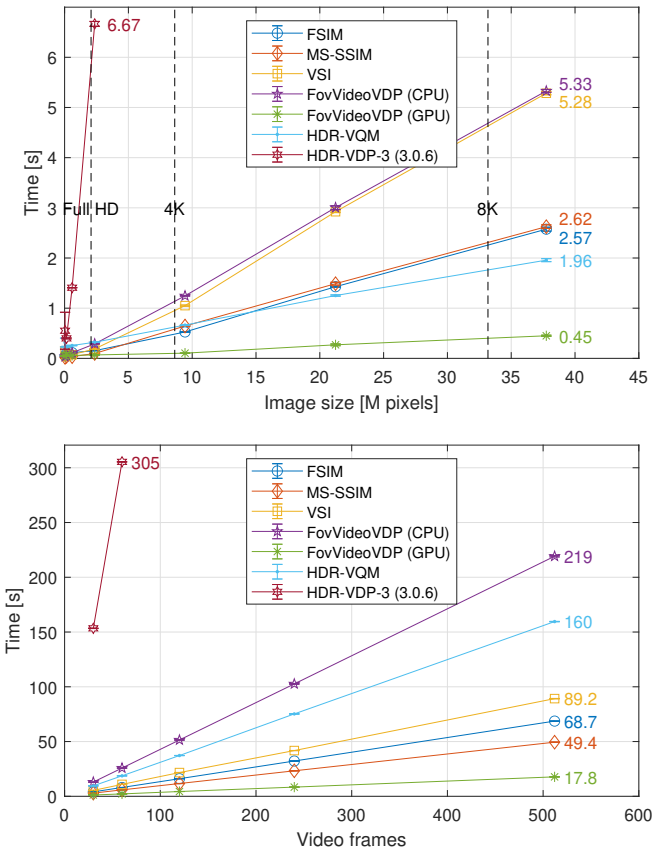

Fig. 2. This figure shows a comparison of the execution time of our metric to other metrics for images (top) and video (bottom). More details on other quality metrics can be seen in Table 1 of the main text.

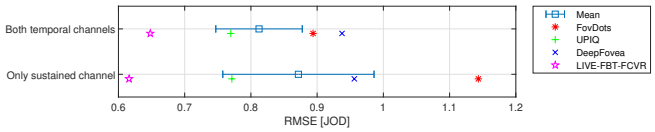

Fig. 3. Prediction error of the variants of the metric that use either both or only a single temporal channel.

5 ABLATIONS AND VARIANTS (CONT.)

Here we describe further ablations and variants of our metric. This complements the *Ablations and variants* section in the main paper.

Temporal channels. One of the key features of our metric is that it separates the video signal into two temporal channels: sustained and transient. We investigated how the prediction performance changes when only the sustained channel is used. The result of that study, shown in Figure 3, indicate that the metric without the transient channel is unable to correctly predict the results of our new FovDots dataset. This shows the importance of FovDots, as it contains artifacts that cannot be found in the other datasets.

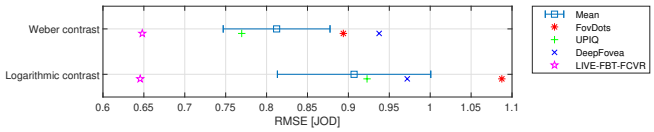

Fig. 4. Prediction error of the variants of the metric that use different representations of contrast.

Contrast representation. Our metric encodes contrast as a Weber fraction: $^{\Delta L}/_{L_a}$, where ΔL represents the amplitude (the value of the coefficient in the Laplacian pyramid) and L_a is the adaptation luminance (refer to Eq.(7) in the main paper). This is different from some other visual metrics, such as VDP and HDR-VDP, which include a photo-receptor non-linearity stage and encode contrast at the early stage of the model. This type of non-linearity transforms luminance into photoreceptor response units, which account for luminance masking (or Weber's law).

One of the simplest forms of photoreceptor non-linearity is the logarithmic function. It can be shown that the logarithmic function is a luminance transducer that is derived from Weber's law; the function that accounts for the fact that visual system is sensitive to ratios of luminance rather than absolute luminance values. We introduced the logarithmic function in the first stage of our metric, before any temporal or spatial filters. This results in contrast coding being simplified to:

$$C_{b,c}(\boldsymbol{x}) = \mathcal{L}'_{b,c}(\boldsymbol{x}) \approx log_{10} \frac{\mathcal{G}_{b,c}(\boldsymbol{x})}{\mathcal{G}_{b+1,c}(\boldsymbol{x})} . \tag{6}$$

$\mathcal{L}'$ represent the Laplacian pyramid of logarithmic values. Note that, as compared to contrast coding in the original version of the metric, there is no need to divide the values by the adaptation luminance (L_a) to account for Weber's law. The right-hand side of the equation shows that this contrast encoding approximates logarithmic contrast.

One challenge of introducing a photoreceptor non-linearity is that it makes the contrast units incompatible with the units used in contrast sensitivity functions. We address this problem by converting the CSF sensitivity to logarithmic contrast:

$$S'_{exfov}(\cdot) = \log_{10}^{-1}\left(\frac{1}{S_{exfov}(\cdot)} + 1\right), \tag{7}$$

where S'_{exfov} represents sensitivity as the inverse of logarithmic contrast.

The results for the metric optimized using our original contrast encoding (Weber contrast) and using logarithmic contrast are shown in Figure 4. A significant drop in performance can be seen when using logarithmic contrast. Since adding this non-linearity also increases the complexity of the metric, we could not justify introducing early logarithmic contrast coding in our metric.

Pooling. Our pooling model involves multiple p-norms, each using a different exponent. The higher exponent values make the metric more sensitive to the largest difference values. Such a "winner-takes-all" strategy is a common pattern found in many mechanisms of the visual system. For example, an exponent between 2 and 3 is found to explain the data on fusion of binocular contrast. However,

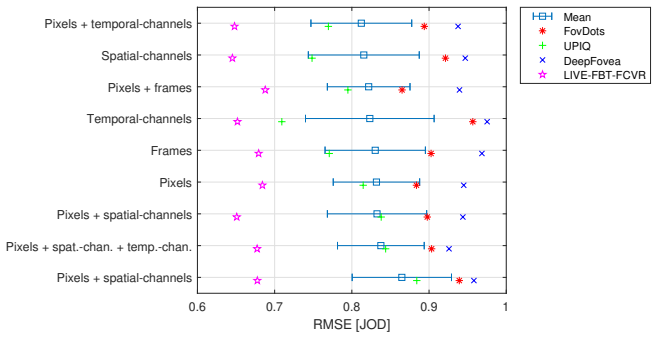

Fig. 5. Prediction error of the variants of the metric that optimized different parameters of the pooling function. The names correspond to the β exponents that were optimized: pixels — $\beta_{\boldsymbol{x}}$; frames — β_f; spatial-channels — β_b; and temporal channels — β_c.

large exponent values also make the metric less stable; metric predictions end up relying mostly on few very high values. For that reason, we restricted the values of all exponents to be less than 3.

When we tried to optimize for all pooling parameters at once (as well as all other metric parameters), we could not achieve good results. For that reason, we optimized one pooling parameter at a time (together with the parameters of the CSF and the masking model) while keeping the remaining exponents equal to 1. When we found the optimum exponent together with a set of other metric parameters, we used those as an initial parameter set for the optimization and added an additional pooling parameter.

The results of optimizing the pooling function, shown in Figure 5, indicate that optimization of a larger number of exponents does not necessarily lead to better performance. The best performance was obtained with optimized $\beta_{\boldsymbol{x}}$ and β_c (controlling the summation of pixels and temporal channels), while the other exponents were set to 1.

6 FOVEATED RENDERING DATASET: ADDITIONAL RESULTS

Here we present the full results from our foveated rendering experiment as described in Section 4 in the main paper.

The results of the pairwise comparison experiments were scaled under Thurstone model V assumptions [Perez-Ortiz and Mantiuk 2017], reducing the comparison rank matrices to linear scales of perceived quality in just-objectionable-difference (JOD) units. 95% confidence intervals were estimated using bootstrapping.

As shown in Figure 7, the sampling percentage s, the temporal anti-aliasing factor β, velocity v, and contrast c all have a noticeable impact on quality. As expected, the higher the sampling percentage, the better the perceived quality; artifacts are also less noticeable on lower values of c. Velocity has an interesting impact on the shapes of the curves, resulting in monotonically decreasing quality (as a function of β) when there is no motion ($v = 0$), to convex or monotonically increasing quality curves for higher velocities. This agrees with the intuition that, for stationary stimuli, temporal artifacts are highly objectionable, while for moving stimuli, the trade-off is non-trivial.

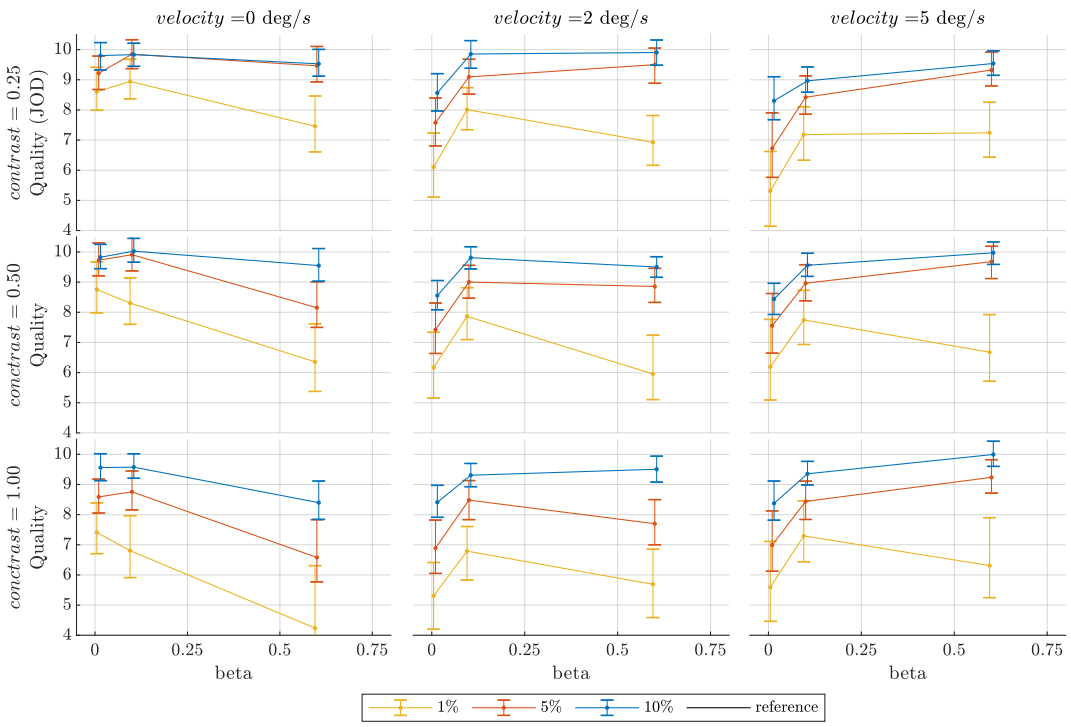

Fig. 6. Results of the psychophysical experiment for $Y = 32.5\,\mathrm{cd/m^2}$ for all pairings of contrast (c) and velocity (v) as a function of the spatio-temporal trade-off factor (β). Colors indicate sampling rate (s). Error bars denote 95% confidence intervals.

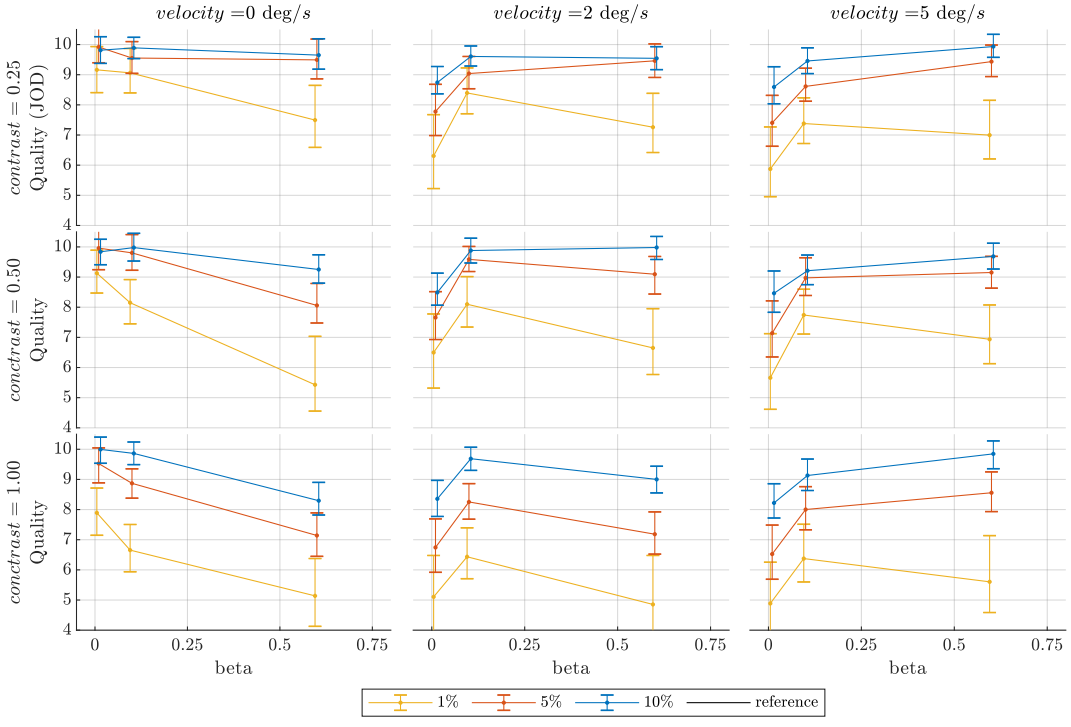

Fig. 7. Results of the psychophysical experiment for $Y = 65\,\mathrm{cd/m^2}$.

Table 1. Result of n-way ANOVA for content and condition parameters of the foveated rendering dataset

	Sum Sq.	Mean Sq.	F	p
velocity	792.96	396.48	569.05	0.00
contrast	1084.83	542.42	778.51	0.00
luminance	0.25	0.25	0.37	0.55
beta	962.21	481.11	690.51	0.00
sampling %	7835.47	3917.74	5622.98	0.00

We performed the experiment on two different mean luminance values: 32.5 cd/m^2 and 65 cd/m^2, both falling within the photopic range of vision. As shown in Figures 6 and 7, the quality curves measured at these different luminance levels show no substantial difference in the shape or magnitude of quality. We speculate that this is due to the limited dynamic range of both the display and our measurements.

To quantitatively analyze the quality curves, we first sampled each quality point assuming normal distributions described by the scaled quality values and their corresponding errors with $N = 35$ samples (identical to the number of observers). N-way ANOVA, as shown in Table 1, confirms the visual analysis, revealing a significant difference with $p << 0.05$ for each free parameter, except for luminance.

7 DATASET MERGING EXPERIMENT

To align the quality scores from DeepFovea and LIVE-FBT-FCVR datasets, we performed a quality matching experiment.

Stimuli. 15 video sequences were selected from each DeepFovea and LIVE-FBT-FCVR datasets, 30 in total. The videos were selected by stratified sampling across the range of subjective quality scores (5 strata) to ensure that the selection covered all quality levels. The reference and test videos were concatenated with a short blank in between so that they could be viewed one after another, always in the same order. The videos were converted to grayscale because (a) it let us simplify the experiment, and (b) the majority of tested metrics ignore color information. A red cross was added at the intended gaze position (center of each video) and the videos were encoded at the *high* quality settings of h265. The videos were assigned random identifiers, which did not reveal the distortion level or any other information.

To provide participants with quality anchors, we also prepared an HTML web page with example images from the UPIQ dataset at the JOD quality levels of 4, 5, .., 10. Only images containing compression or banding artifacts were selected to facilitate matching to the h264/h265 artifacts found in the two video datasets.

Participants. 8 expert participants were recruited for this experiment. Due to COVID-19 restrictions, they were asked to complete the experiment at home, on a display that had a diagonal size of 24" or more. They were provided with a table of viewing distances corresponding to different screen dimensions.

Experimental procedure. Each expert was asked to view videos enlarged to the full screen size, to keep the viewing distance and to keep the gaze on the fixation cross. They were instructed that those videos are intended for foveated viewing and they were all familiar with the concept. They were asked to rate each video using a JOD scale, including fractional numbers if the video fell between two JOD levels shown in the training web page Answer were submitted using an online form.

Results. We excluded from the results the data from a single expert, who consistently rated all videos at a much lower JOD than the rest, likely because they failed to maintain proper fixation on the target throughout the experiment.The mean JOD ratings of select videos for both datasets are shown in Figure 8. The plots show that the relation between the native quality scores of each dataset (DMOS or MOS) and JODs can be explained by a linear regression.

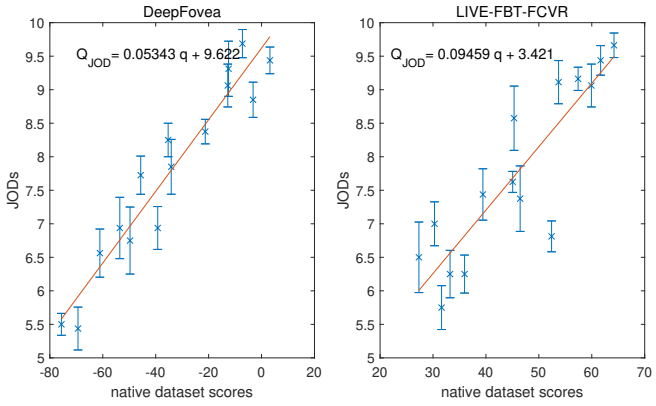

Fig. 8. This figure shows the results of the dataset merging experiment. Crosses with error bars indicate the rating of stimuli from the DeepFovea or LIVE-FBT-FCVR datasets using the JOD scale. Error bars indicate standard error.

REFERENCES

S.J. Daly. 1993. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, Andrew B. Watson (Ed.). Vol. 1666. MIT Press, 179–206. https://doi.org/10.1117/12.135952

J. M. Foley. 1994. Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A* (1994).

Gordon E. Legge and John M. Foley. 1980. Contrast masking in human vision. *JOSA* 70, 12 (dec 1980), 1458–71.

Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. FovVideoVDP : A visible difference predictor for wide field-of-view video. *ACM Transaction on Graphics* (2021).

Maria Perez-Ortiz and Rafal K. Mantiuk. 2017. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint* (dec 2017). arXiv:1712.03686 http://arxiv.org/abs/1712.03686

AB Watson and JA Solomon. 1997. Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A* 14, 9 (1997), 2379–2391.

Sophie Wuerger, Maliha Ashraf, Minjung Kim, Jasna Martinovic, María Pérez-Ortiz, and Rafal K. Mantiuk. 2020. Spatio-chromatic contrast sensitivity under mesopic and photopic light levels. *Journal of Vision* 20, 4 (apr 2020), 23. https://doi.org/10.1167/jov.20.4.23