Supplementary of ColorVideoVDP: A visual difference predictor for image, video and display distortions

RAFAŁ K. MANTIUK, University of Cambridge, UK
PARAM HANJI, University of Cambridge, UK
MALIHA ASHRAF, University of Cambridge, UK
YUTA ASANO, Reality Labs, USA
ALEXANDRE CHAPIRO, Reality Labs, USA

ACM Reference Format:
Rafał K. Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. 2024. Supplementary of ColorVideoVDP: A visual difference predictor for image, video and display distortions. In *SIGGRAPH 2024 Technical Papers (SIGGRAPH '24 Technical Papers), Jul 28– Aug 1, 2024, Denver, USA.* ACM, New York, NY, USA, Article 129, 9 pages. https://doi.org/10.1145/3658144

This supplementary document describes:

- the procedure used to convert from the cone contrast units used in castleCSF to the DKL contrast used in ColorVideoVDP (Sec. 1);
- the contrast encoding (Sec. 2) and contrast masking functions (Sec. 3) that were considered for ColorVideoVDP and are included in the ablations;
- several mathematical techniques that were necessary to make the model differentiable (Sec. 4);
- timings of ColorVideoVDP compared to VMAF (Sec. 5);
- the experiment used to obtain a JOD quality scaling for the LIVEHDR dataset (Sec. 6);
- details on the content used for XR-DAVID dataset (Sec. 7).

1 CONTRAST SENSITIVITY IN THE DKL COLOR SPACE

The castleCSF [Ashraf et al. 2024] model predicts sensitivity in different contrast units than those used by ColorVideoVDP. To use castleCSF in ColorVideoVDP, we need to convert between the two contrast units. The contrast conversion procedure is similar to the one used in [Kim et al. 2021].

The sensitivity used in castleCSF is defined as the inverse of cone contrast:

$$S_{\mathrm{cc}} = \sqrt{3}\left(\left(\frac{\Delta\mathfrak{L}}{\mathfrak{L}_0}\right)^2 + \left(\frac{\Delta\mathfrak{M}}{\mathfrak{M}_0}\right)^2 + \left(\frac{\Delta\mathfrak{S}}{\mathfrak{S}_0}\right)^2\right)^{-0.5}, \tag{1}$$

where $\Delta\mathfrak{L}$, $\Delta\mathfrak{M}$, and $\Delta\mathfrak{S}$ are the amplitudes of the cone responses for the stimuli and $\mathfrak{L}_0$, $\mathfrak{M}_0$, and $\mathfrak{S}_0$ are the cone responses for the corresponding background. ColorVideoVDP encodes contrast for the three cardinal dimensions of the DKL space ($c \in \{\mathrm{Ach}, \mathrm{RG}, \mathrm{YV}\}$) as:

$$C_c = \frac{\mathcal{L}_c}{L}, \tag{2}$$

where $L = \mathfrak{L}_0 + \mathfrak{M}_0$ is luminance, and $\mathcal{L}_c$ is the amplitude of the achromatic, red-green, or yellow-violet response in the DKL space (an increment in that space and also the coefficient of the Laplacian pyramid). Here, we omitted unnecessary indices from Eq. (6) in the main paper. The sensitivity in the DKL-contrast units is the inverse of Eq. (2):

$$S_{\mathrm{DKL},c} = C_c^{-1} = \frac{L}{\mathcal{L}_c}, \tag{3}$$

To convert S_{cc} into $S_{\mathrm{DKL},c}$, we first obtain from castleCSF the sensitivity S_{cc} along each cardinal color direction (c) of the DKL color space. Then, we find the contrast C_c that results in the threshold cone contrast corresponding to the predicted sensitivity (S_{cc}^{-1} since the threshold contrast is the inverse of the sensitivity). As there is no closed-form solution, C_c needs to be found by non-linear root finding. In the optimization loop, we allow the DKL contrast C_c to change only along the given cardinal color direction. This calculation is repeated for each spatial frequency, luminance, and color direction. The sensitivities in the DKL contrast space are precomputed and stored in a look-up table for later use by ColorVideoVDP, as explained in Sec. 3.4 of the main paper.

2 CONTRAST ENCODING

Multiplicative contrast normalization. Once we have separated the band-limited contrast in each frequency band and visual channel (two achromatic and two chromatic), we want to encode it in a way that correlates well with the perceived magnitude of the contrast. To this end, many visual models employ the normalization by the contrast sensitivity function $S(\cdot)$

$$C' = C\,S(\rho_b, L_{\mathrm{bkg}}, c), \tag{4}$$

where C is the physical and C' is encoded contrast, ρ_b is the peak frequency of the band b, c is the channel index, and L_{bkg} is the background luminance. Since contrast can be expressed as $C = \Delta L / L_{\mathrm{bkg}}$, we can write:

$$C' = C\,S(\rho_b, L_{\mathrm{bkg}}, c) = \frac{\Delta L}{L_{\mathrm{bkg}}}\frac{L_{\mathrm{bkg}}}{\Delta L_{\mathrm{thr}}} = \frac{\Delta L}{\Delta L_{\mathrm{thr}}}, \tag{5}$$

where ΔL_{thr} is the luminance difference corresponding to the detection threshold. It follows that C' encodes multiples of the detection

Authors' addresses: Rafał K. Mantiuk, rafal.mantiuk@cl.cam.ac.uk, University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge, UK, CB3 0FD; Param Hanji, param.hanji@gmail.com, University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge, UK, CB3 0FD; Maliha Ashraf, ma905@cam.ac.uk, University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge, UK, CB3 0FD; Yuta Asano, yasano@meta.com, Reality Labs, , Redmond, USA, ; Alexandre Chapiro, alex@chapiro.net, Reality Labs, , Sunnyvale, USA, ;

SIGGRAPH '24 Technical Papers, July 28–Aug 1, 2024, Denver, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/3658144

threshold. Notably, it is equal to 1 when the contrast C is exactly at the threshold value. This property is necessary for any contrast encoding employed by our method because masking models, discussed in the next section, rely on it. The encoding is plotted in the left panel of Fig. 1.

The above contrast normalization by sensitivity brings several benefits. Daly [1993] showed that such a normalization is necessary to unify masking predictions across spatial frequencies. Peli et al. [1991] demonstrated that this normalization accounts for contrast matching across luminance levels.

Additive contrast normalization. The multiplicative normalization from Eq. (5) can be justified by the physical limits of the eye's optics at high frequencies, or by the lateral inhibition at low frequencies [Barten 1999]. However, it does not explain contrast constancy across spatial frequencies [Georgeson and Sullivan 1975] — that is, the observation that the perceived magnitude of large suprathreshold contrast appears the same regardless of spatial frequency. Contrast constancy can be better explained by Kulikowski's model of contrast matching [Kulikowski 1976], in which the detection threshold is subtracted from (rather than divided by) the absolute contrast

$$C'_K = C - \frac{1}{S(\rho_b, L_{\text{bkg}}, c)}\,. \tag{6}$$

Note that the inverse of the contrast sensitivity in this equation is the threshold visibility contrast. Kulikowski [1976] demonstrated that perceived contrast is matched across luminance levels when the corresponding C'_K is matching (the sensitivity S in the equation varies across luminance levels). Kulikowski's model assumes that the detection thresholds are caused by additive (neural) noise, and therefore, the visual system "eliminates" the noise by subtracting it from the contrast signal.

To support both contrast polarities, we modify the above formula as follows:

$$C' = \text{sgn}(C)\,\max\left\{g\left(|C| - \frac{1}{S(\rho_b, L_{\text{bkg}}, c)}\right) + 1, 0\right\}\,, \tag{7}$$

where g is the gain that controls the strength of the contrast above the threshold. As in the previous case, the constant 1 is used to ensure that the encoded contrast has a value of 1 at the threshold. This is an important modification as it allows us to represent contrasts below the detection threshold. The sign function ensures that we preserve the polarity of the contrast. The encoding is plotted in the right panel of Fig. 1.

To check how well each contrast encoding represents perceived contrast, in the following sections we generate contrast-matching predictions and compare them with those reported in the literature.

2.1 Matching chromatic and achromatic contrast

Since our metric needs to evaluate the impact of both achromatic and chromatic distortions, we need to ensure that the magnitude of both is correctly matched. Switkes and Crognale [1999] measured color matches along multiple directions in the color space, including the cardinal directions of the DKL space that we use on ColorVideoVDP. They found that suprathreshold contrast can be matched across achromatic and chromatic dimensions by simple multiplicative scaling. In Fig. 2, we plot their contrast-matching

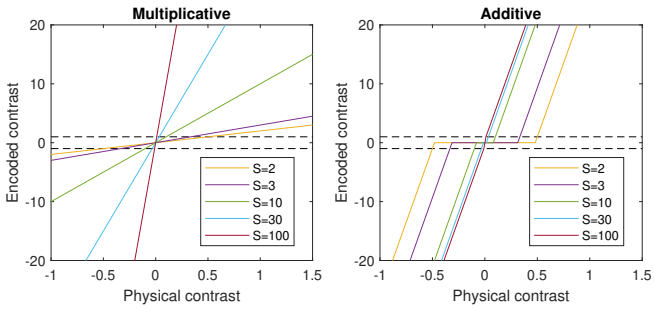

Fig. 1. *Mapping from input physical contrast to encoded contrast for multiplicative (left) and additive (right) encoding.*

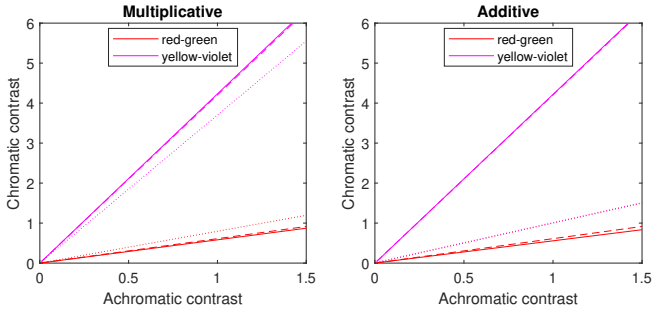

Fig. 2. *The lines connecting the perceived magnitude of achromatic contrast that matches in appearance the perceived magnitude of chromatic contrast, either red-green or yellow-violet. The dashed lines represent the model based on the measurements of Switkes and Crognale [1999]. The dotted lines show the predictions of the original models from Eq.* (5) *and Eq.* (7) *and the continuous lines the same models but after the adjustment. Note that in the right plot for additive contrast encoding the two dotted lines overlap each other.*

model as two dashed lines — matching achromatic contrast to either reg-green or yellow-violet chromatic directions. The dotted lines, representing the predictions of our contrast encoding models from Eq. (5) and Eq. (7). The dotted lines in the left plot show that the multiplicative encoding with the normalization by the CSF is close to the measurements of Switkes and Crognale, but with some inaccuracy that grows with contrast. This is because the CSF is measured for very small (threshold) contrasts, and any inaccuracy at such fine scales is amplified for large contrast values. The additive model, shown as two overlapping dotted lines in the right plot, cannot predict contrast matches across color directions.

To improve the accuracy of these matches, we introduced corrections into the contrast encoding equation for the multiplicative contrast encoding:

$$C' = m_c\, C\, S(\rho_b, L_{\text{bkg}}, c)\,, \tag{8}$$

where $m_c = \begin{bmatrix}1 & 1 & 1.45 & 0.95\end{bmatrix}$ for the color channels:

$$c \in \{\text{Ach}\mathcal{S}, \text{Ach}\mathcal{T}, \text{RG}, \text{YV}\} \tag{9}$$

corresponding to the achromatic sustained, achromatic transient, chromatic red-green, and chromatic yellow-violet channels. The

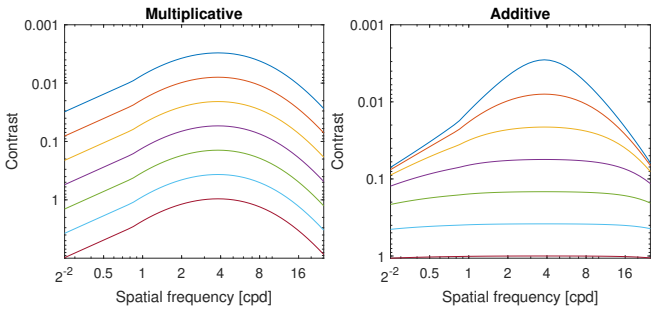

Fig. 3. *The lines connecting the matching magnitude of contrast across frequencies, as predicted by the multiplicative (left) and additive (right) contrast encoding.*

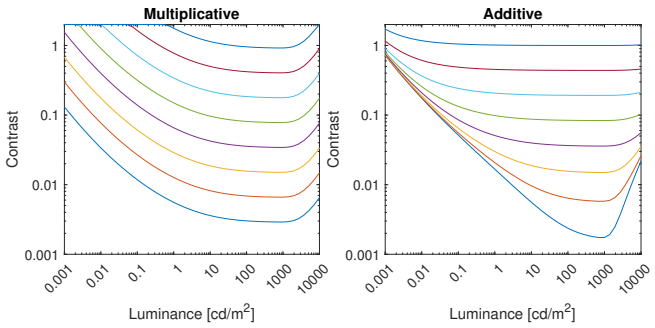

Fig. 4. *The lines connecting the matching magnitude of contrast across luminance, as predicted by the multiplicative (left) and additive (right) contrast encoding.*

additive contrast encoding needs to include an additional gain factor:

$$C' = \mathrm{sgn}(C)\max\left\{g\, m_c\left(|C| - \frac{1}{S(\rho_b, L_{\mathrm{bkg}}, c)}\right) + 1, 0\right\}, \qquad (10)$$

where $m_c = \begin{bmatrix}1 & 1 & 1.7 & 0.237\end{bmatrix}$. The adjusted contrast encoding equations are plotted as continuous lines Fig. 2.

2.2 Matching contrast across frequencies

The seminal paper of Georgeson and Sullivan [1975] demonstrated contrast constancy — the observation that the perceived magnitude of contrast differs across spatial frequencies when the contrast is small (near the detection threshold) but when the contrast is far above the threshold, there is little difference in the perceived magnitude of contrast. This property is better captured by the additive contrast encoding model, shown in the right panel of Fig. 3. This contrast constant effect is easy to observe in everyday life — objects seen from close and far distances will fall into very different spatial frequency ranges. Yet, we do not observe changes in object appearance as we move closer or further away from them. The multiplicative contrast encoding technique does not directly model the property of contrast constancy.

2.3 Matching contrast across luminance

Multiple authors investigated whether contrast constancy generalizes across luminance levels [Georgeson and Sullivan 1975; Hess 1990; Kulikowski 1976; Peli 1995; Peli et al. 1991] and they all observed quite a significant deviation from contrast constancy, especially when the luminance drops significantly below photopic levels. However, there is no agreement on how to model such a deviation from contrast constancy. Both Kulikowski [1976] and Georgeson [1991] proposed additive contrast encoding to explain their contrast matching data. However, Peli [1995] demonstrated that a multiplicative model better explains the data when contrast is seen naturally rather than using a dichoptic presentation used in other studies (each eye sees a different luminance). The predictions for both models are shown in Fig. 4.

3 CONTRAST MASKING

The main purpose of the masking model is to transform physical differences in contrast between two images or frames into perceived differences. Here, consider three models of masking: a model based on contrast transducer functions, such as the one proposed by Watson and Solomon [Watson and Solomon 1997], a model based on mutual masking, such as the one proposed in Daly's original VDP [Daly 1993], and the similarity formula, used in SSIM and other metrics.

Contrast transducer. The difference between two band-limited images in Watson and Solomon's [Watson and Solomon 1997] masking model is expressed as a difference between two contrast transducer functions:

$$D_{b,c,f}(\boldsymbol{x}) = \left|t(C'^{\mathrm{test}}_{b,c,f}(\boldsymbol{x})) - t(C'^{\mathrm{ref}}_{b,c,f}(\boldsymbol{x}))\right| \qquad (11)$$

where b is the index of the spatial frequency band, c is the channel index (two achromatic and two chromatic), and f is the frame index. $C'^{\mathrm{test}}_{b,c,f}$ and $C'^{\mathrm{ref}}_{b,c,f}$ correspond to the encoded contrast in the test and reference images (refer to the main paper). The transducer function is formulated as:

$$t(C'_{b,c,f}) = g_{\mathrm{T}}\frac{\mathrm{sgn}(C'_{b,c,f}(\boldsymbol{x}))\left|C'_{b,c,f}(\boldsymbol{x})\right|^p}{0.2 + \sum_i k_{i,c}\left(\left|C'_{b,i,f}\right|^{q_c} * g_{\sigma_{\mathrm{sp}}}\right)(\boldsymbol{x})}, \qquad (12)$$

where p and q_c are the parameters of the model. The masking parameter q_c is set separately for each channel (two achromatic and two chromatic channels). g_{T} is the gain of the transducer that lets us control the range of visual difference values. The constant of 0.2 was selected so that the facilitation (the dip in the contrast discrimination function) coincides with the measurements, as explained in Sec. 3.2. The expression in the denominator pools contrast in a local spatial neighborhood by convolving with a Gaussian kernel $g_{\sigma_{\mathrm{sp}}}$ with the standard deviation of σ_{sp}. The sum in this expression pools contrast across channels according to the cross-masking coefficient $k_{c,i}$. C' is the encoded contrast, as explained in Sec. 2.

Mutual masking. Alternatively, the difference between two band-limited images can be expressed using the mutual masking model, as proposed by Daly [Daly 1993]:

$$D_{b,c,f}(\boldsymbol{x}) = \frac{\left|C'^{\mathrm{test}}_{b,c,f}(\boldsymbol{x}) - C'^{\mathrm{ref}}_{b,c,f}(\boldsymbol{x})\right|^p}{1 + (C^{\mathrm{mask}}_{b,c,f}(\boldsymbol{x}))^{q_c}} \qquad (13)$$

First, the mutual masking of test and reference bands is calculated as [Daly 1993, p.192]:

$$C^{\mathrm{mm}}_{b,c,f}(\boldsymbol{x}) = \min\left\{\left|C'^{\,\mathrm{test}}_{b,c,f}(\boldsymbol{x})\right|, \left|C'^{,\mathrm{ref}}_{b,c,f}(\boldsymbol{x})\right|\right\}. \tag{14}$$

Then, similarly to the previous model, the mutual masking signal is pooled in a small local neighborhood by convolving with a Gaussian kernel $g_{\sigma_{\mathrm{sp}}}$, and combined across channels, accounting for cross-channel masking:

$$C^{\mathrm{mask}}_{b,c,f}(\boldsymbol{x}) = \sum_i k_{i,c}(C^{\mathrm{mm}}_{b,i,f} * g_{\sigma_{\mathrm{sp}}})(\boldsymbol{x}). \tag{15}$$

One shortcoming of the mutual masking model is that it does not account for self-masking — the case in which the reference contrast is 0 and the test contrast is attenuated when its value is high. In practice, $D_{b,c,f}$ can reach very high values (over 10 000) when the sensitivity is high. This is unrealistic behavior as neurons cannot encode values of such high dynamic ranges. For that reason, we need to limit the range of contrast difference values with a smooth clamping function:

$$\hat{D}_{B,c,f}(\boldsymbol{x}) = \frac{D_{\max} D_{B,c,f}(\boldsymbol{x})}{D_{\max} + D_{B,c,f}(\boldsymbol{x})}, \tag{16}$$

where $D_{\max}$ is the maximum value that the visual difference can attain.

Similarity. The alternative masking model, found in SSIM and many other metrics, can be formulated as:

$$D_{b,c,f}(\boldsymbol{x}) = D_{\max} - D_{\max}\frac{2\left|C'^{\,\mathrm{test}}_{b,c,f}(\boldsymbol{x})\right|\left|C'^{\,\mathrm{ref}}_{b,c,f}(\boldsymbol{x})\right| + \epsilon}{\left(\hat{C}^{\mathrm{test}}_{b,c,f}\right)^2(\boldsymbol{x}) + \left(\hat{C}^{\mathrm{ref}}_{b,c,f}\right)^2(\boldsymbol{x}) + \epsilon}, \tag{17}$$

where $\epsilon = D_{\max} - 1$ is selected so that the resulting visual difference is 1 when the test contrast is at the detection threshold ($C'^{\,\mathrm{test}}_{b,c,f} = 1$) and the reference contrast is 0. $D_{\max}$ controls the maximum value of the visual difference. The standard formula typically uses the same values in both nominator and denominator. Here, we modify the denominator so that it contains masking signal $\hat{C}^{\mathrm{test}}_{b,c,f}$ and $\hat{C}^{\mathrm{ref}}_{b,c,f}$ associated with the test and reference images. For the test image

$$\hat{C}^{\mathrm{test}}_{b,c,f}(\boldsymbol{x}) = \sum_i k_{i,c}(C'^{\mathrm{test}}_{b,i,f} * g_{\sigma_{\mathrm{sp}}})(\boldsymbol{x}). \tag{18}$$

and the masking signal for the reference frame is computed analogously.

3.1 Self-masking

To give more insights into the three masking models explained above and the two contrast encodings, we plot the model predictions for the case of self-masking in Fig. 5. Self-masking is the case in which the contrast (in the test) is masked by itself and is not influenced by the contrast in the reference image (the reference contrast is). This scenario may appear e.g. when the reference image is a uniform field, and the test image contains a pattern that we want to detect. All the plots in Fig. 5 show that the visual difference is equal to 1 (the dashed horizontal line) when the test contrast is at the detection threshold (equal to $1/S$). The additive contrast encoding (right column) will result in small contrast being ignored when the sensitivity is low, while the multiplicative encoding will compress contrast below the detection threshold. The mutual masking and similarity models are remarkably similar to each other in terms of self-masking.

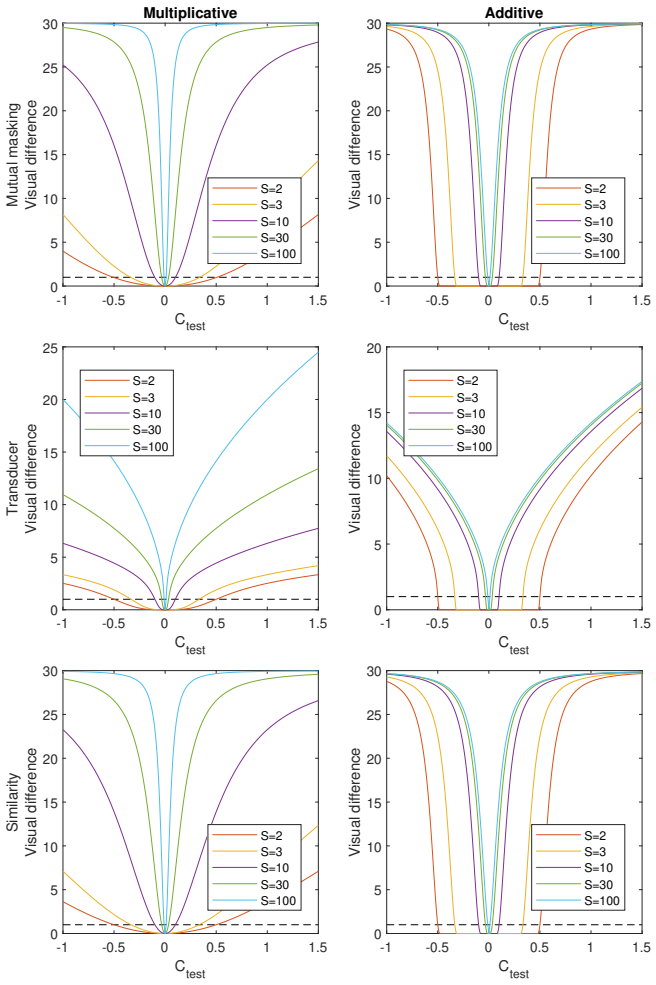

Fig. 5. *Model response for self-masking — when the reference contrast is 0 (is a uniform field) but the test contrast (x-axis) and sensitivity (colors) vary. The predictions are shown for all combinations of contrast coding (columns) and masking models (rows). The dashed horizontal line at $D = 1$ represents the difference at the detection threshold.*

3.2 Contrast masking

The standard contrast masking experiment involves showing a test pattern superimposed on top of a masking pattern, such as the one shown in Fig. 7. We simulate the same scenario for our six combinations of masking models and contrast encodings and show the results in Fig. 6. In each plot, we include the measurements of contrast masking from [Foley 1994] as red dots. We expect the contour lines to follow the curve formed by those measurements. In the plots, we can see that only the transducer models the facilitation that makes patterns easier to detect when the masker is near the

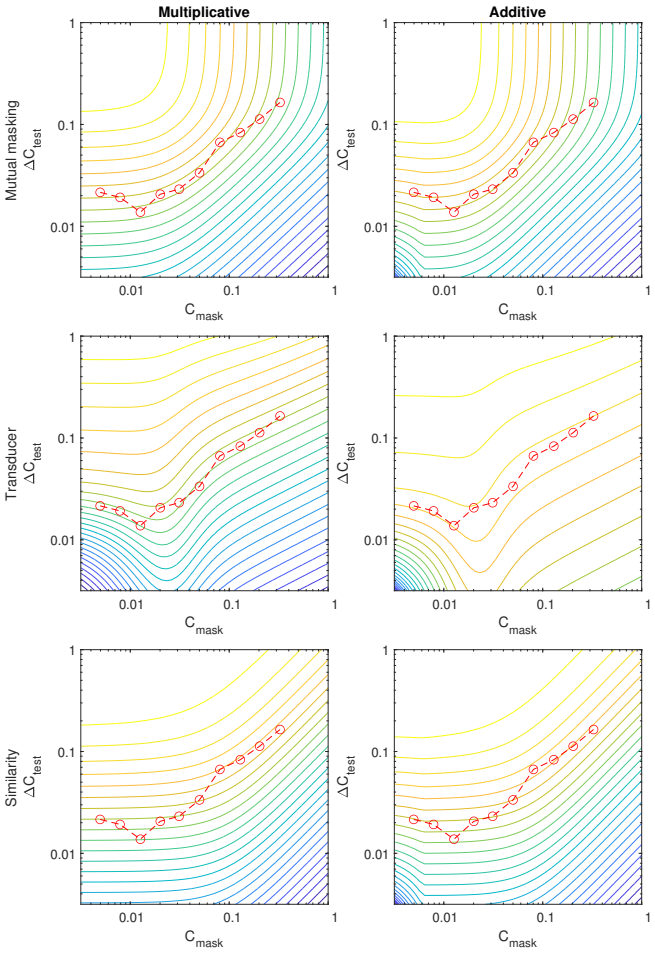

Fig. 6. *Model response for contrast masking. The reference image contains a certain physical contrast C^{mask} (x-axis), and the test image contains the physical contrast of $C^{mask} + \Delta C^{test}$ (y-axis). The contour lines denote the response of the model (blue for the smallest and yellow for the largest values). The red circles denote the contrast masking measurements from [Foley 1994, Figure 3b, data for KMF]. The axes roughly correspond to those found in the example in Fig. 7.*

detection threshold. This is shown as a small dip in the measurements by Foley [1994]. Mutual masking and similarity models show similar behavior.

4 TRAINING CONSIDERATIONS

While the contrast encoding and masking models proposed in the previous section are well-defined for an arbitrary contrast, these models are not differentiable out-of-the-box and, therefore, cannot be used in gradient-based optimization. To make them differentiable, we replace the sign function with a hyperbolic tangent function:

$$\mathrm{sgn}(C) \approx \tanh(10000\, C)\,. \tag{19}$$

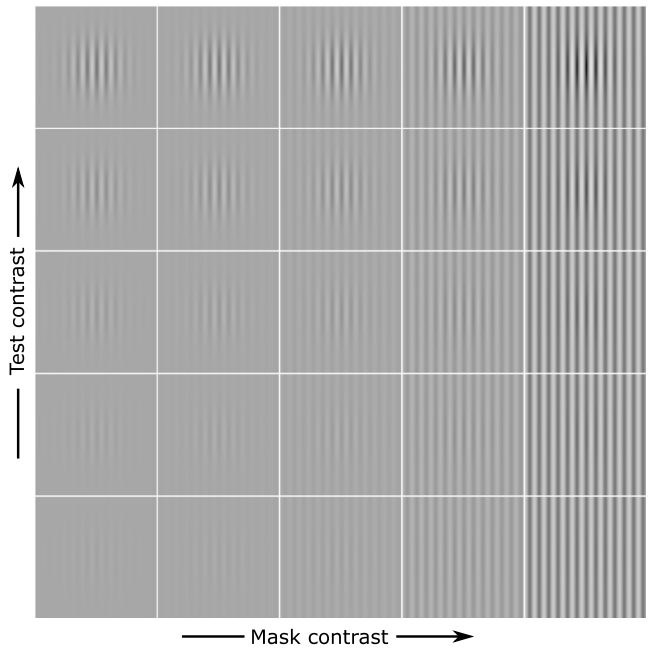

Fig. 7. *The example of typical stimuli used in contrast masking experiments. A test Gabor patch is superimposed (added) on the background of a sinusoidal grating of the same frequency (a masker). As the contrast of the masker is increased (towards the right), a higher test contrast of the Gabor patch is needed to detect it.*

The exponential function is not differentiable at 0, therefore, we need to approximate it as:

$$|C|^p \approx (|C| + \epsilon)^p - \epsilon^p\,, \tag{20}$$

where $\epsilon = 0.00001$.

5 QUALITY METRIC TIMINGS

We measured the times required to process 50 video frames at resolutions ranging from 720p to 4K. The measurements were averaged across 5 runs and excluded the times required to load and decode the frames. The measured times, illustrated in Fig. 8, show that ColorVideoVDP processes large videos faster than VMAF (note however, that VMAF was running on a CPU) and in about the same time as FovVideoVDP. It must be noted, however, that unlike ColorVideoVDPboth VMAF and FovVideoVDP operate only on luma or luminance and ignore color channels. VMAF features are much less expensive to extract and FovVideoVDP processes only 2 visual channels (sustained and transient achromatic) compared to the 4 channels processed by ColorVideoVDP.

6 LIVE HDR DATASET ALIGNMENT EXPERIMENT

In order to enable training on multiple datasets, it is essential to bring their quality scores to a common scale. This ensures that quality scores from one dataset are directly comparable and equivalent to those from another dataset. While both XR-DAVID and UPIQ employ the same quality units (JODs), LIVE HDR was originally collected using the mean-opinion-score (MOS) units. To account for

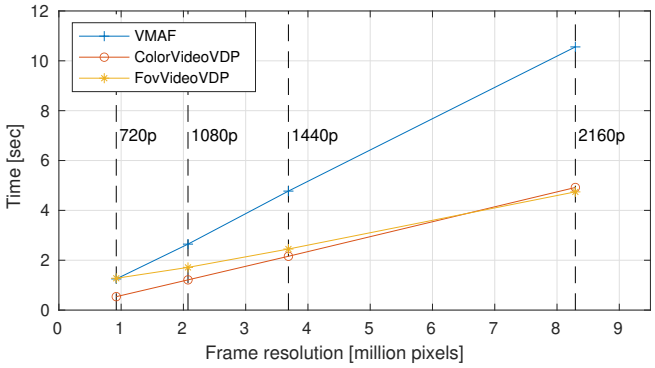

Fig. 8. *The processing times for 50-frame video of different resolutions (x-axis). ColorVideoVDP and FovVideoVDP were run on an Nvidia Quadro RTX 8000 GPU, while VMAF was run on 4 cores of an Intel Xeon Gold 5218 CPU @ 2.30 GHz. The dashed vertical lines indicate the standard video resolutions, from 720p to 2160p (4K). All the reported times were averaged across five runs.*

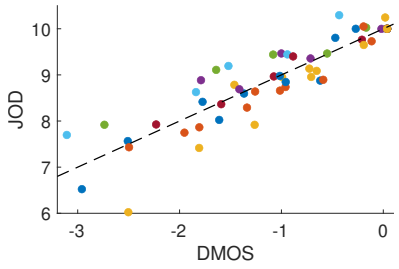

Fig. 9. *We conducted a subjective study, obtaining* JOD *scores for a subset of videos from the LIVE HDR dataset. A linear fit was obtained as follows:* $\text{JOD} = -0.054 * \text{DMOS} + 0.037$*, which was deemed acceptable.* JOD *values are shifted from 0 to 10 by convention.*

this discrepancy, we ran an additional experiment that let us scale LIVE HDR MOS values in the JOD units.

We first calculated differences of mean opinion scores (DMOS) by subtracting the score of the reference (highest quality video) from each condition. Next, in order to maximize the sampling of quality across the dataset, we carefully selected 10 videos that contained a wide distribution of MOS values. For each of these videos, we then selected 6 different levels of distortion. We used this content to conduct an experiment using the same experimental procedure as for XR-DAVID, as explained in the main paper. Display and environment settings were adapted to mimic those used in the original LIVE HDR study. 12 participants took part in the study, and completed 10 repetitions (batched) of *ASAP* sampling, for a total of $10 \times 6 \times 10 = 600$ trials each. The resulting JOD values for each condition were used to find a linear mapping from the LIVE HDR DMOS scores to our experimentally derived JOD scores. The resulting linear fit, illustrated in Fig. 9, demonstrates a satisfactory level of accuracy. Thus, we applied this mapping to the entire LIVE HDR dataset, effectively rescaling it to JOD units.

Table 1. XR-DAVID assets

Video	Source	Duration
Bonfire	Kindel Media [link]	5.6s
Business	Kindel Media [link]	5.6s
Caminandes	Blender Foundation [link]	4.8s
Couple	RDNE Stock project [link]	5.6s
Dance	Anna Shvets [link]	5.6s
Emojis	emirkhan bal [link]	5.6s
Foliage	German Korb [link]	5.6s
Icons	Generated by the authors	5.6s
Panel	Generated by the authors	5.6s
Cellphone	Lina Fresco [link]	5.6s
River	Theresa. Nguyen [link]	3.7s
Snow	SwissHumanity Stories [link]	5.6s
River	Theresa. Nguyen [link]	3.7s
VR	Generated by the authors	5.6s
Wiki	Generated by the authors	5.2s

7 XR-DAVID CONTENT

Thumbnails of select scenes used in the XR-DAVID dataset are shown in Fig. 10. The source data for each video is shown in Table 1. Note that for longer videos, sections of approximately 5 seconds from the start of each video were used in the study, with the exception of Caminandes where a shot from approximately 1:54-2:00 of *"Caminandes 3: Llamigos"* was used. Videos were downsampled to a resolution of 910×540 using bilinear interpolation, which produces an effective visual resolution of $\approx$ 40ppd when seen by users from a distance of 28.9'. Prior to display, videos were upscaled using nearest-neighbor interpolation by a factor of 2 to 1920×1080 to match the native resolution of the display. Detailed numerical results of our experiment are shown in Fig. 11 for each of the 3 levels of intensity per artifact studied.

REFERENCES

Maliha Ashraf, Rafał K. Mantiuk, Alexandre Chapiro, and Sophie Wuerger. 2024. castleCSF — A Contrast Sensitivity Function of Color, Area, Spatio-Temporal frequency, Luminance and Eccentricity. *Journal of Vision* 24 (2024), 5. https://doi.org/10.1167/jov.24.4.5

Peter G. J. Barten. 1999. *Contrast sensitivity of the human eye and its effects on image quality.* SPIE Press. 208 pages. http://books.google.com/books?hl=en&lr=&id=kPyyBAomC4cC&pgis=1

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

John M. Foley. 1994. Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A* 11, 6 (jun 1994), 1710. https://doi.org/10.1364/JOSAA.11.001710

B Y M A Georgeson and G D Sullivan. 1975. Contrast constancy: deblurring in human vision by spatial frequency channels. *The Journal of Physiology* 252, 3 (1975), 627–656.

MA Georgeson. 1991. Contrast overconstancy. *Journal of the Optical Society of America A* (1991). http://www.opticsinfobase.org/josaa/ViewMedia.cfm?id=4026&seq=0

R. F. Hess. 1990. Vision at low light levels: role of spatial, temporal and contrast filters*. *Ophthalmic and Physiological Optics* 10, 4 (Oct. 1990), 351–359. https://doi.org/10.1111/j.1475-1313.1990.tb00881.x

Minjung Kim, Maryam Azimi, and Rafał K. Mantiuk. 2021. Color Threshold Functions: Application of Contrast Sensitivity Functions in Standard and High Dynamic Range Color Spaces. *Electronic Imaging* 33, 11 (jan 2021), 153–1–153–7. https://doi.org/10.

2352/ISSN.2470-1173.2021.11.HVEI-153

J.J. Kulikowski. 1976. Effective contrast constancy and linearity of contrast sensation. *Vision Research* 16, 12 (jan 1976), 1419–1431. https://doi.org/10.1016/0042-6989(76)90161-9

Eli Peli. 1995. Suprathreshold contrast perception across differences in mean luminance: effects of stimulus size, dichoptic presentation, and length of adaptation. *Journal of the Optical Society of America A* 12, 5 (may 1995), 817. https://doi.org/10.1364/JOSAA.12.000817

Eli Peli, Jian Yang, Robert Goldstein, and Adam Reeves. 1991. Effect of luminance on suprathreshold contrast perception. *Journal of the Optical Society of America A* 8, 8 (aug 1991), 1352. https://doi.org/10.1364/JOSAA.8.001352

Eugene Switkes and Michael A. Crognale. 1999. Comparison of color and luminance contrast: apples versus oranges? *Vision Research* 39, 10 (may 1999), 1823–1831. https://doi.org/10.1016/S0042-6989(98)00219-3

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. http://www.opticsinfobase.org/abstract.cfm?URI=josaa-14-9-2379

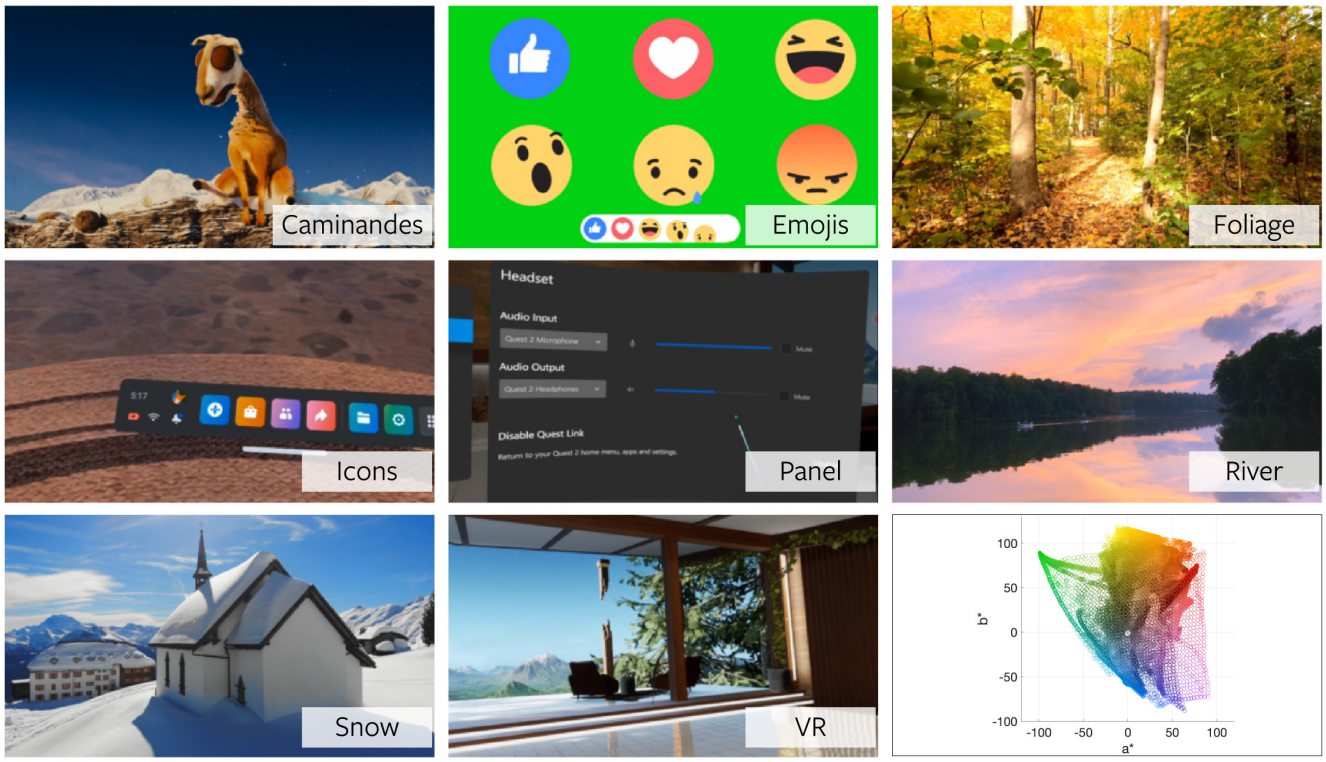

Fig. 10. *Thumbnails of 8 out of 14 scenes used in the XR-DAVID study. The bottom-right plot shows the color gamut coverage of all the pixels in the reference videos on a CIELAB a* b* plane. The thumbnails of the videos containing human subjects could not be included due to institutional policy.*

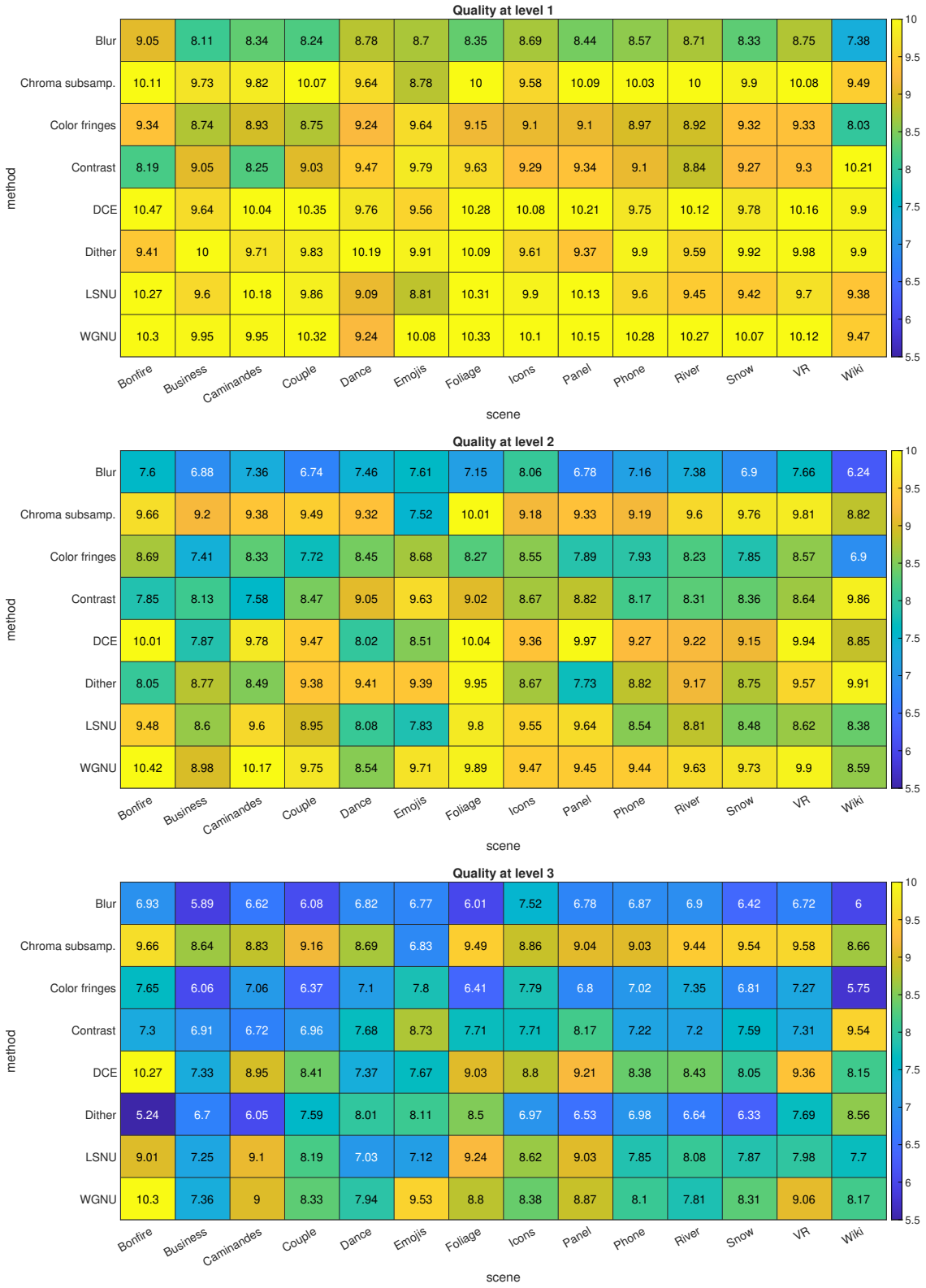

Fig. 11. *The results of our XR-DAVID color video quality experiment. For each base video, 8 different artifacts were studied at 3 intensity levels (1:top, 2:middle, and 3:bottom). The responses are scaled on a single perceptual* JOD *scale, counting down from 10 by convention. Increasing magnitudes of perceived distortion can be observed at stronger distortion levels (top-to-bottom). In addition, large differences in artifact visibility can be observed across content (columns).*