ColorVideoVDP-ML: Visual Difference Predictor with a neural regressor for image and video quality prediction

Dounia Hammou, Fei Yin, Rafał K. Mantiuk
University of Cambridge, UK
{dh706, fy277, rafal.mantiuk}@cl.cam.ac.uk

***Abstract*—ColorVideoVDP-ML is a full-reference image and video quality metric, which combines a visual difference predictor with a machine learning regressor. It employs ColorVideoVDP as an encoder of visual information and visible differences, and a neural architecture as a regressor of the final quality scores. It benefits from low-level vision models of contrast sensitivity and masking, found in ColorVideoVDP, as well as its ability to handle both SDR and HDR content, and to adapt to arbitrary viewing conditions, display resolution, and size. Two neural regressors, one based on MLPs and one on Transformers, are trained on multiple video and image quality datasets, all scaled in the Just-Objectionable-Difference (JOD) units. The regressors let us account for higher-level factors influencing perceived image quality, such as saliency or texture differences, and provide a way to fine-tune and specialize the metric in handling a specific type of distortions. The code for the metric can be found at https://github.com/gfxdisp/ColorVideoVDP.**

***Index Terms*—Visual difference predictor, image and video quality assessment, neural regressor, saliency, attention**

I. Introduction

ColorVideoVDP [1] is a full-reference quality metric for images and videos, which accounts for the physical specifications of a display (emitted color, size, resolution) and which models low-level human spatio-temporal and chromatic vision. Because it incorporates contrast sensitivity (castleCSF [2]) and contrast masking models, it can predict whether introduced distortions are visible to the human eye. However, since it relies on a simple pooling of visible contrast differences, it cannot account for higher-level factors that affect image quality, such as saliency or texture similarity. For example, if we have a video with a salient foreground attracting all the attention, ColorVideoVDP will still penalize equally the distortions in the foreground and background. Since background can cover a larger area, the metric can skew its predictions towards less important parts of the frame. Another problematic case for ColorVideoVDP is image parts containing high-entropy (random) textures, such as foliage or grass. For a human observer, a slight difference in such a texture, such as rotation or displacement, makes no difference. However, since ColorVideoVDP compares local differences between the test and reference contrast, it will label those as containing large distortions.

To address those shortcomings, we introduce ColorVideoVDP-ML — ColorVideoVDP with machine learning regressors, which can account for high-level factors and improve the regression of quality values. We use the original ColorVideoVDP to encode visual information as perceived contrast of the test, reference image, and also as visible difference between the two, which accounts for contrast masking. To reduce such visual information to a manageable size, we compute local statistics for patches over a spatial region of one visual degree and then pass those to one of the deep-learning architectures: either a saliency-based or a transformed-based regression network, which predict the final quality score.

We could train ColorVideoVDP-ML both effectively and efficiently with three training strategies. First, the extracted features with local statistics reduced the size of the training data from terabytes of raw video files to a few gigabytes of feature vectors, which could be efficiently used for training. Second, we were able to train and validate the metric on multiple datasets, all scaled on the Just-Objectionable-Difference (JOD) units. Because JOD units are generally consistent across datasets, we could train on multiple datasets at the same time, and monitor for overfitting using other datasets. Finally, the videos used to train one of the models were augmented by extracting smaller videos, trimmed in the temporal dimension, and cropped in the spatial dimension. We assumed that smaller segments of a video have the same quality score as the entire video. This lets us train larger models and also increases the amount and diversity of the training data.

II. Related works

We briefly review visual difference predictors, as those are the most relevant to the proposed metric. The original difference predictor [3] was proposed in the 90s as a way to incorporate psychophysical models of low-level vision into the assessment of image quality. It included models of contrast sensitivity (dependence of sensitivity on spatial frequency) and contrast masking (the reduced visibility of differences in the presence of a masking signal, such as textures). HDR-VDP-1 [4] introduced modeling of glare and luminance masking to account for the visibility of distortions in high-dynamic-range images. HDR-VDP-2 [5] was the main redesign of the original VDP, which was calibrated on multiple psychophysical datasets. It was also the first VDP metric that could regress visual differences, typically represented as a difference map, into single-value quality scores that were well

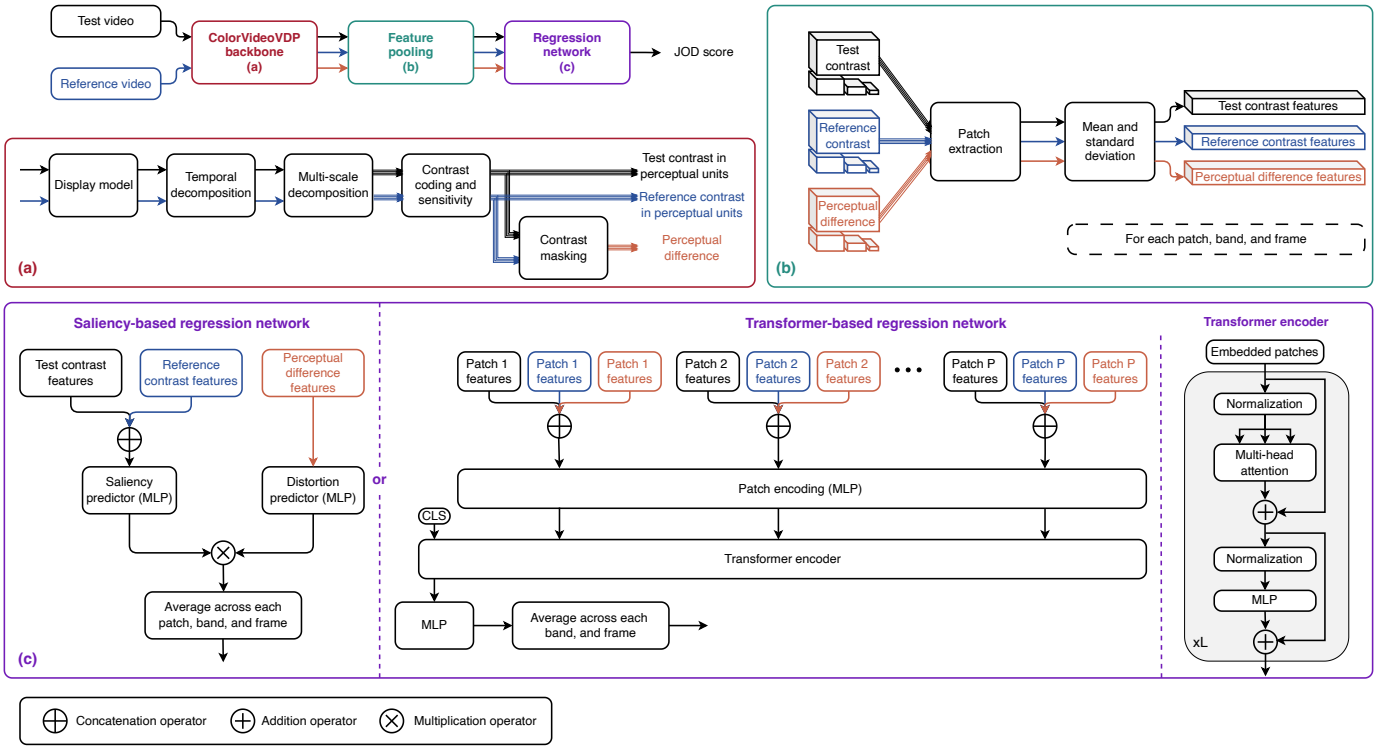

Fig. 1. Architecture overview of the proposed ColorVideoVDP-ML. (a) provides details on the ColorVideoVDP metric that is used as a backbone. (b) explains how we pooled the features from the contrast information extracted from ColorVideoVDP. (c) showcases the two architectures that we employed to regress the features.

correlated with mean opinion scores [6]. HDR-VDP-3 [7] further improved quality and introduced variants of the predictor (e.g., side-by-side vs. flicker presentation of compared images). FovVideoVDP [8] was another major redesign, which simplified many components of HDR-VDP-3 to make it more suitable for video quality predictions. It introduced temporal channels, which modeled the visibility of temporal distortions, such as flicker. FovVideoVDP could also model foveation, which is the reduction of sensitivity to distortion outside the foveated (gazed) region. ColorVideoVDP [1], which our metric is based on, removed the modeling of foveation in favor of modeling color vision, both in the spatial and temporal domains. ColorVideoVDP relies on a modern spatiotemporal and chromatic contrast sensitivity function, castleCSF [2].

All reviewed VDPs share the same weakness — inability to model higher-level factors affecting image quality. Our ColorVideoVDP-ML is an attempt to address this weakness.

III. ColorVideoVDP-ML

In this section, we introduce ColorVideoVDP-ML. The objective of this metric is to leverage the components of visual difference predictors based on the psychophysical foundations of the low human visual system and extend them to capture higher-level neural processing through the use of neural network architectures. An overview of the metric pipeline is presented in Figure 1, with a detailed explanation provided in the subsequent sections.

A. Perceptual difference features

The backbone of our metric is ColorVideoVDP [1] — a recent visual difference predictor for image, video, and display distortions. It is used to encode test and reference images and their difference into perceptual representations. At a high level, as shown in Figure 1-(a), ColorVideoVDP first models the display by transforming input content from its native color space (e.g., BT.2020 with the PQ EOTF) to a linear color space that represents absolute quantities of light emitted from a display. This step is responsible for handling both SDR and HDR content and for accounting for the physical characteristics of the display (size, resolution, viewing distance). Following that, temporal filters are used to decompose videos into four perceptually meaningful channels: achromatic, red-green, yellow-violet, and transient channel. Those are then decomposed into a multi-scale bandpass pyramid (the Laplacian pyramid). Each band is then modulated by the corresponding contrast sensitivity (based on castleCSF [2]) and passed to a contrast masking model, which predicts the visibility of the distortions. In the original ColorVideoVDP, such contrast differences are simply pooled across the spatial dimension, bands, channels, and frames. Instead, we pass this information to the machine learning regressor.

B. Feature pooling

The main difficulty of training a video quality metric is the gigantic amount of raw video data used to predict relatively few labels. Video datasets can easily take several terabytes

when stored in uncompressed format. To reduce the amount of data to a manageable size, we pool the ColorVideoVDP output in non-overlapping spatial patches, as shown in Figure 1-(b). Each patch has a width and height of one visual degree at the highest frequency band. At lower frequency bands, which are subsampled in the pyramid, the number of pooled coefficients remains the same, so that the patch increases in the size measured in visual degrees. The pooling computes six statistics for each patch: mean and standard deviation of absolute values of the test and reference contrast (modulated by the contrast sensitivity), and mean and standard deviation of the perceptual differences. Those statistics are computed separately for the four visual channels, for each band, and each frame. Since images lack a transient achromatic channel, the statistics of that channel are set to 0 for images.

C. Regression network architecture

We regress the final quality scores using either one of the two neural architectures, explained below.

1) ColorVideoVDP-ML-Saliency: The first architecture attempts to learn the saliency of different image regions and weigh the impact of the local distortions accordingly. As illustrated in Figure 1-(c), the saliency-based regression network utilizes one MLP network ϕ_D to regress visible differences (D) into quality scores, and another MLP network ϕ_S to compute the weight for the first network prediction. This allows us to modulate the influence of the distortions found in each patch by the saliency computed from the test (T) and reference (R) patch statistics.

The final quality score is computed as:

$$Q_{\text{JOD}} = 10 - k \sum_{f=1}^{F} \sum_{b=1}^{B} \sum_{p=1}^{P} \mathbb{R}\left(\phi_S\left(T_{f,b,p} \oplus R_{f,b,p}\right)\right) \cdot \mathbb{R}\left(\phi_D\left(D_{f,b,p}\right)\right) \tag{1}$$

where $k = \frac{1}{FBP}$ is the normalization constant and the summation is performed across patches p, bands b and frames f. $\mathbb{R}$ denotes the ReLU activation function that constrains the saliency and distortion scores to remain non-negative, as our goal is to model only perceptual degradation rather than enhancement. The symbol $\oplus$ denotes the concatenation operator. Vectors T, R and D contain 8 values — two statistics computed for four channels.

While we do not train our saliency predictor ϕ_S on eye-tracking data, we believe that using this joint-training of both predictors on quality datasets, the model would learn to assign higher scores for salient regions (as shown later in Section IV-D).

2) ColorVideoVDP-ML-Transformer: While the previous architecture incorporates semantic information to weigh the distortions' visibility, it does not account for the relationship between the different patches. Therefore, as an alternative regressor, we use an architecture based on Transformers [9].

As shown in Figure 1-(c), the metric can leverage the Transformer network architecture to capture inter-patch dependencies and their impact on the global perceived quality. Given each patch p, we concatenate the corresponding test T, reference R, and difference D features and feed them to the patch encoding module ϕ_P, implemented as a linear layer. The resulting patch embeddings, concatenated with a learnable classification token CLS, are then fed to the Transformer encoder to predict the distortion score for each band and frame, which were then aggregated across all bands and frames, as described previously in Eq. (1).

As illustrated in the figure, we omit positional embeddings, as they did not yield any performance improvement. Additionally, we explored incorporating the features across all frequency bands, by including a "band" embedding as in [10]. However, considering that contrast information is similar across frequency bands and to reduce the complexity of the model, we opted not to include it.

D. Implementation details

For the ColorVideoVDP-ML-Saliency model, the saliency predictor is implemented as an MLP with 4 layers and 48 hidden units, along with a dropout rate of 0.2. The distortion predictor consists of 3 layers with 24 hidden units and the same dropout configuration. Batch normalization is intentionally omitted from both predictors to preserve the units of the ColorVideoVDP features, which are expressed in meaningful perceptual units. Both networks utilize ReLU activation functions.

In the ColorVideoVDP-ML-Transformer variant, the patch encoding module is implemented as a single linear layer that transforms the input features of size 24 to a token of dimension 256. The Transformer encoder consists of 4 layers, each with 8 attention heads. Within the encoder's MLP blocks, we adopt the GeLU activation function, following standard practice in Transformer architectures. The regression MLP head is implemented as a single normalization layer, followed by a linear layer, and a ReLU activation function.

Both architectures include two additional learnable parameters: the "baseband weight" and the "image weight". The baseband weight is used to scale the model prediction for the baseband, following the approach in ColorVideoVDP, ensuring they are normalized to be compatible with the magnitude of higher frequency bands. The image weight is introduced to adapt the model for image quality assessment, where no temporal information is available. It adjusts the resulting quality scores to be consistent with those predicted for video content.

The number of parameters for each architecture, their inference time, as well as the learnable baseband and image weight, are reported in Table I.

IV. Experiments

A. Training and testing datasets

We utilized a total of eight datasets, four used for both training and validation (with an 80/20 train/test split with no shared scenes) and the remaining four were reserved solely for validation. We used a large number of validation datasets to monitor training and prevent overfitting. Details of all datasets

TABLE I
THE MODEL PARAMETER NUMBER, SIZE, INFERENCE TIME, AND TRAINABLE PARAMETERS OF COLORVIDEOVDP AND BOTH COLORVIDEOVDP-ML NEURAL ARCHITECTURES. THE INFERENCE TIME IS MEASURED FOR A 60 FRAMES 4K HDR10 VIDEO USING AN NVIDIA GEFORCE RTX 4090 GPU.

Model	Number of parameters	Model size [Mb]	Inference time [ms/frame]	baseband weight	image weight
ColorVideoVDP	33	0.001	85.65	-	-
ColorVideoVDP-ML-Saliency	9.4K	0.15	87.84	2.12	0.65
ColorVideoVDP-ML-Transformer	3.2M	37	91.03	0.94	0.90

are provided in Table II. All training datasets' scores were scaled in the Just-Objectionable-Difference (JOD) units [11], which provided consistent quality scores across the datasets. Moreover, the MOS scores of the LIVE HDR dataset were converted to JOD scores by linear mapping, as described in the supplementary of [1]. While the conversation may introduce some inaccuracy, we found that training including this dataset resulted in higher performance on the testing datasets.

B. Training methodology

To allow efficient training, the pooled features (statistics stored T, R, and D tensors) were extracted once at the beginning of training, cached, and reused across all training epochs. Despite this optimization, training the Transformer-based model remained memory-intensive. To mitigate this, we used data augmentation when training ColorVideoVDP-ML-Transformer. We applied random spatial cropping to a range between 25% and 75% of the original resolution and sampled temporal sequences of random durations between 0.5 and 1 second. While we acknowledge that the perceptual quality of a spatio-temporal patch may not perfectly reflect that of the full video [19], we believe that spatial cropping facilitates the Transformer's ability to learn inter-patch spatial relationships more effectively.

For optimization, we employ the Adam optimizer [20], with a learning rate of 0.001 for the saliency-based model and 0.0001 for the Transformer-based model. Training is conducted over 250 epochs, and the final model is selected based on the epoch that achieves the best overall performance across all validation datasets. Our framework is implemented using the PyTorch framework, and training is performed on an NVIDIA GeForce RTX 4090 GPU.

C. Metric performance

We compare the performance, measured as the Spearman correlation, of our proposed ColorVideoVDP-ML with eight state-of-the-art full-reference image and video quality metrics [7], [21]–[26] on the testing datasets. Metrics that do not natively account for HDR content were adapted using the PU21 encoding [27]. Furthermore, metrics that do not natively account for the viewing distance (introduced in the HDR-VDC dataset) were adapted using SAST rescaling [28].

The results in Figure 2 show that the challenge dataset (HDRSDR-VQA) proved difficult for many metrics, with popular metrics, such as LPIPS, PSNR-Y, HDR-VDP-3 and

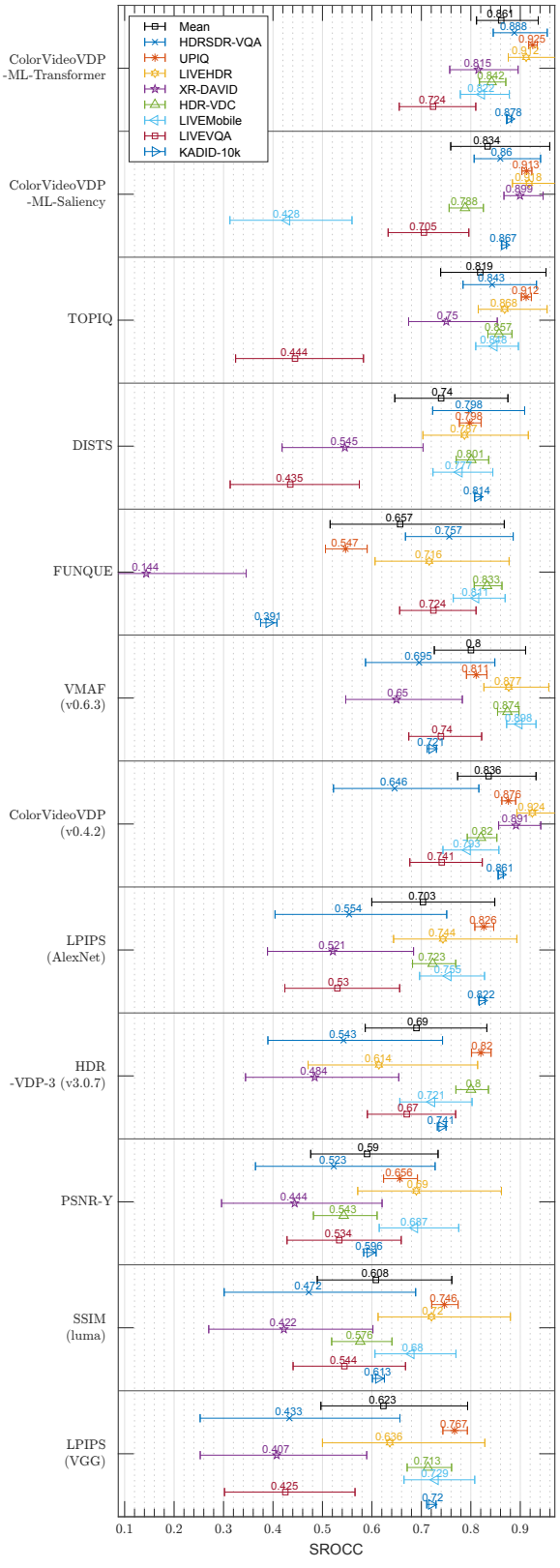

Fig. 2. Spearman correlation performance of ColorVideoVDP-ML compared with existing quality metrics. The performance is reported on the testing datasets and the test portion of the datasets employed for both training and testing. The error bars denote 95% confidence intervals. The metrics are sorted by their performance on the HDRSDR-VQA (challenge) dataset. The results of TOPIQ on the Kadid-10K dataset were removed as it was trained on it.

TABLE II
DATASETS USED FOR TRAINING AND TESTING.

Dataset	Used for	Type	Scenes	Conditions	Distortions
HDRSDR-VQA [12] (challenge dataset)	Train & test	SDR/HDR videos	20	360	H.265, bicubic upscaling
UPIQ [13]	Train & test	SDR/HDR images	84	4159	34 distortion types
LIVE HDR [14]	Train & test	HDR video	21	210	H.265, bicubic upscaling
XR-DAVID [1]	Train & test	SDR video	14	336	8 display artifacts
HDR-VDC [15]	Test	HDR video	16	464	AV1, lanczos upscaling, varying viewing distance and display luminance level
LIVE VQA [16]	Test	HDR video	10	150	H.264, MPEG-2, transmission
LIVE Mobile VQA [17]	Test	HDR video	10	200	H.264, wireless packet loss, frame freezes, temporally varying compression rates
KADID-10k [18]	Test	SDR image	81	10125	25 distortion types

SSIM, achieving correlations close to 0.5. ColorVideoVDP and VMAF also struggled — closer inspection showed that those metrics underpredicted the quality of high-resolution videos and overpredicted that of low-resolution videos. TOPIQ and DISTS achieved the best performance on this dataset, despite showing weaker performance on some other datasets. Both variants of our metrics outperformed other metrics on the testing portion of HDRSDR-VQA, but more importantly, they retained good performance on the testing datasets, confirming that they were not overfitted. ColorVideoVDP already performed very well on the other training datasets — LIVEHDR, XR-DAVID, and UPIQ. Both variants of our metric achieved similar performance on those datasets.

When we consider testing datasets, we can observe both gains in performance with respect to ColorVideoVDP (ColorVideoVDP-ML-Transformer on HDR-VDC and LIVE-Mobile), but also losses in performance (ColorVideoVDP-ML-Saliency also on LIVEMobile). Both variants of our metric perform worse than VMAF [25] and FUNQUE [26] on the LIVE VQA dataset. This is an older dataset, which includes transmission error distortions that were missing in the training datasets. Overall, ColorVideoVDP-ML-Transformer performs better in those tests than ColorVideoVDP-ML-Saliency, however, it requires 3 orders of magnitude more parameters (see Table I).

D. Qualitative analysis: Distortion visibility map visualization

To assess the effectiveness of our model in incorporating semantic information during quality prediction, we visualize the distortion visibility maps generated by the two variants of the ColorVideoVDP-ML model in Figure 3. For comparison, we also include a baseline version of ColorVideoVDP-ML trained without any semantic information, using only a simple distortion predictor trained solely on perceptual difference features.

As shown in the figure, the baseline model — lacking semantic awareness — assigns relatively uniform importance across the entire image and tends to emphasize background distortions that are unlikely to influence perceived quality. The saliency-based variant partially mitigates this issue by reducing attention to the background and focusing more on the central object, such as the woman seated on the bench. Notably, the Transformer-based model demonstrates the greatest improvement, concentrating attention primarily on the woman while largely disregarding background regions. This indicates a more human-aligned perception of quality.

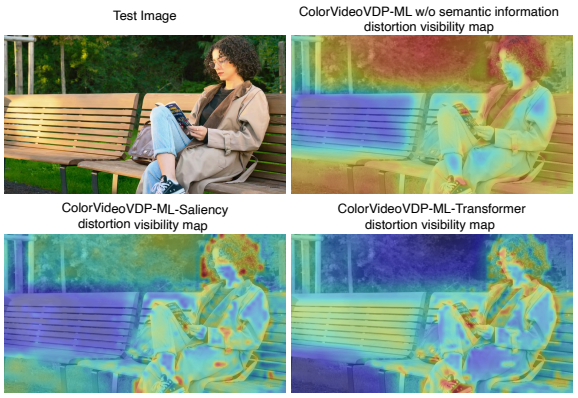

Fig. 3. Distortion visibility maps of a frame extracted from the HDRSDR-VQA challenge dataset using three different models: A baseline ColorVideoVDP-ML with no semantic information, ColorVideoVDP-ML-Saliency, and ColorVideoVDP-ML-Transformer.

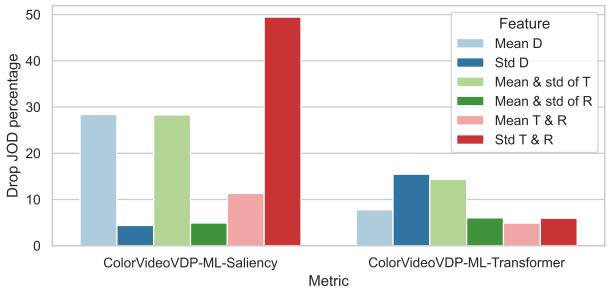

Fig. 4. Feature sensitivity analysis of the ColorVideoVDP-ML metrics. The bar plot illustrates the percentage change of the JOD quality prediction after setting to 0 the statistic mentioned in the legend. The values are computed for the test portion of the HDRSDR-VQA dataset.

E. Feature sensitivity analysis

To understand which features our model relies on the most, we perform a feature sensitivity analysis. We set the values of selected statistics to 0 and then measure how much the quality prediction changes. The results are reported in Figure 4.

The figure shows that both variants of the ColorVideoVDP-ML rely on the test contrast features more than the reference contrast features to derive the semantic information, which aligns with previous findings [29]. Moreover, we can observe

that the transformer-based network relies more on the perceptual difference features. In contrast, the saliency-based network assigns more importance to the test and reference features in comparison to the perceptual difference features. This can be due to the similar importance we are giving to the distortion and saliency predictors, which can lead the network to slightly overfit on the test and reference features.

V. Conclusion

In this paper, we introduced ColorVideoVDP-ML, a full-reference image and video quality assessment metric that integrates low-level human visual system modeling with high-level neural processing. Inspired by the success of both classical visual difference predictors and deep learning-based approaches, our method unifies both paradigms into a single, generalizable metric. We proposed two neural regression architectures: a saliency-based MLP that learns to weight spatial regions based on their perceptual importance, and a Transformer-based model that captures spatial relationships across regions. Trained on four diverse image and video quality datasets and evaluated on eight, our metric achieves state-of-the-art performance, particularly on challenging datasets such as the HDRSDR-VQA challenge dataset.

References

[1] Rafal K. Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro, "ColorVideoVDP: A visual difference predictor for image, video and display distortions," *ACM Transactions on Graphics*, vol. 43, no. 4, July 2024.

[2] Maliha Ashraf, Rafał K. Mantiuk, Alexandre Chapiro, and Sophie Wuerger, "castleCSF — A contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity," *Journal of Vision*, vol. 24, no. 4, pp. 5–5, 04 2024.

[3] S.J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, Andrew B. Watson, Ed., vol. 1666, pp. 179–206. MIT Press, 1993.

[4] Rafal Mantiuk, Scott J. Daly, Karol Myszkowski, and Hans-Peter Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," San Jose, CA, Mar. 2005, p. 204.

[5] Rafał K. Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 1–14, July 2011.

[6] Manish Narwaria, Rafal K. Mantiuk, Mattheiu Perreira Da Silva, and Patrick Le Callet, "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1, pp. 010501, Jan. 2015.

[7] Rafal K. Mantiuk, Dounia Hammou, and Param Hanji, "HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content," , no. arXiv:2304.13625, Apr. 2023, arXiv:2304.13625 [cs, eess].

[8] Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney, "FovVideoVDP: A visible difference predictor for wide field-of-view video," *ACM Transaction on Graphics*, vol. 40, no. 4, pp. 49, 2021, Citation Key: Mantiuk2021.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[10] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang, "MUSIQ: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5148–5157.

[11] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Transactions on Image Processing*, vol. 29, pp. 1139–1151, 2019.

[12] Bowen Chen, Cheng han Lee, Yixu Chen, Zaixi Shang, Hai Wei, and Alan C. Bovik, "HDRSDR-VQA: A subjective video quality dataset for HDR and SDR comparative evaluation," 2025.

[13] Aliaksei Mikhailiuk, María Pérez-Ortiz, Dingcheng Yue, Wilson Suen, and Rafał K Mantiuk, "Consolidated dataset and metrics for high-dynamic-range image quality," *IEEE Transactions on Multimedia*, vol. 24, pp. 2125–2138, 2021.

[14] Zaixi Shang, Joshua P Ebenezer, Alan C Bovik, Yongjun Wu, Hai Wei, and Sriram Sethuraman, "Subjective assessment of high dynamic range videos under different ambient conditions," in *2022 IEEE international conference on image processing (ICIP)*. IEEE, 2022, pp. 786–790.

[15] Dounia Hammou, Lukáš Krasula, Christos G Bampis, Zhi Li, and Rafał K Mantiuk, "The effect of viewing distance and display peak luminance—HDR AV1 video streaming quality dataset," in *International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2024, pp. 193–199.

[16] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[17] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[18] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, "KADID-10k: A large-scale artificially distorted iqa database," in *International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[19] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik, "Patch-VQ:'Patching Up' the video quality problem," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14019–14029.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[23] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[24] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, "TOPIQ: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, 2024.

[25] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock, "VMAF: The journey continues," *Netflix Technology Blog*, vol. 25, no. 1, 2018.

[26] Abhinau K Venkataramanan, Cosmin Stejerean, and Alan C Bovik, "FUNQUE: Fusion of unified quality evaluators," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2147–2151.

[27] Rafal K. Mantiuk and Maryam Azimi, "PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[28] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, "Self-adaptive scale transform for iqa metric," in *IEEE international symposium on circuits and systems (ISCAS)*. IEEE, 2013, pp. 2365–2368.

[29] Wei Zhang and Hantao Liu, "Study of saliency in objective video quality assessment," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1275–1288, 2017.