

# FovVideoVDP: A visible difference predictor for wide field-of-view video

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

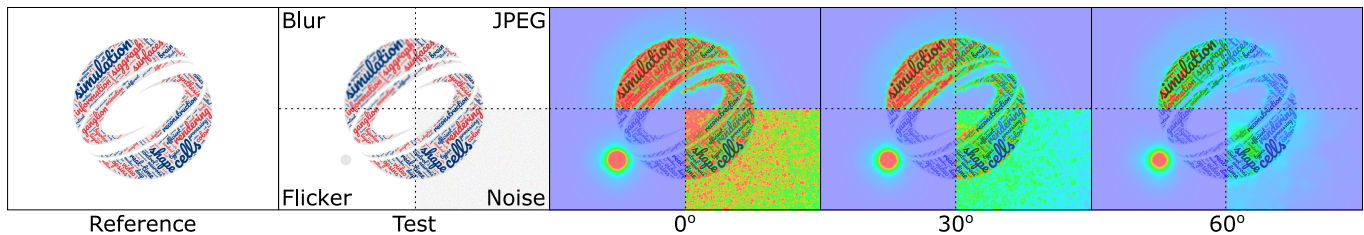


Fig. 1. The heat-maps show our metric’s predictions for 4 types of distortions (blur, JPEG compression, 30 Hz flicker and Gaussian additive noise) at three different eccentricities. The flicker is simulated as a dot that appears in every second frame. All types of artifacts are predicted to be much less noticeable when seen with peripheral vision at large eccentricities. Refer to Figure 20 for the color scale of the heat-map.

FovVideoVDP is a video difference metric that models the spatial, temporal, and peripheral aspects of perception. While many other metrics are available, our work provides the first practical treatment of these three central aspects of vision simultaneously. The complex interplay between spatial and temporal sensitivity across retinal locations is especially important for displays that cover a large field-of-view, such as Virtual and Augmented Reality displays, and associated methods, such as foveated rendering. Our metric is derived from psychophysical studies of the early visual system, which model spatio-temporal contrast sensitivity, cortical magnification and contrast masking. It accounts for physical specification of the display (luminance, size, resolution) and viewing distance. To validate the metric, we collected a novel foveated rendering dataset which captures quality degradation due to sampling and reconstruction. To demonstrate our algorithm’s generality, we test it on 3 independent foveated video datasets, and on a large image quality dataset, achieving the best performance across all datasets when compared to the state-of-the-art.

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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2021 Association for Computing Machinery.

0730-0301/2021/8-ART49 \$15.00

<https://doi.org/10.1145/3450626.3459831>

CCS Concepts: • **Computing methodologies** → **Perception**.

Additional Key Words and Phrases: VDP, video quality, foveated rendering, perceptual metric

## 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. *ACM Trans. Graph.* 40, 4, Article 49 (August 2021), 19 pages. <https://doi.org/10.1145/3450626.3459831>

## 1 INTRODUCTION

Quality metrics are a basic feature of modern image processing algorithms. Whether by avoiding costly user studies, providing the user with a standardized measure of the expected effect, or flagging areas of content where artifacts are present, metrics are widely relied upon. Furthermore, accurate metrics are essential to the development of cost functions — a central component of optimization processes and highly popular machine learning methods.

Image and video difference metrics have historically been the subject of a great deal of investigation. In spite of this, the complexity of the human visual system does not allow for a simple algorithmic representation. While visual computing researchers have leveraged discoveries in vision science to craft ever more accurate and biologically inspired metrics, many modern applications remain poorly served by available metrics that do not account for certain aspects of novel displays, resulting in sub-optimal performance. This is particularly relevant in rendering applications for wide-field-of-view (FOV) displays, notably virtual and augmented reality (VR/AR).

In the central (foveal) region, the human eye can perceive frequencies in excess of 60 pixels per visual degree. Matching such a

high spatial resolution uniformly across a wide FOV display, such as a 110° VR headset, demands unreasonable performance from algorithms and hardware. This issue can be addressed by employing *foveated rendering*, i.e., rendering pixels more densely in the central (foveal) region, and more sparsely in the periphery of vision, away from the gaze location. For such methods to be effective, a full resolution frame must be perceived as indistinguishable from traditional, full-resolution rendering. Unfortunately, most state-of-the-art quality metrics are not suitable to estimate the quality of foveated content, as they do not consider all the necessary aspects of vision. Crucially, most practical metrics ignore the significant effect of peripheral perception. Despite the loss of spatial acuity with eccentricity (angular distance from the center of fixation), spatio-temporal artifacts can become even more prominent in wide FOV displays as peripheral vision retains a high degree of sensitivity to temporal changes [Hartmann et al. 1979].

This work presents the first unified full-reference metric of visible differences over space, time, and eccentricity<sup>1</sup>. Our metric is based on psychophysical models of human vision and is rigorously validated on both existing data sets and the results of a novel psychophysical study. The new study has been designed to characterize how artifacts produced by foveated rendering methods affect quality.

## 2 RELATED WORK

We begin by discussing existing metrics, followed by the applications that motivate metrics like ours, in particular foveated rendering. We discuss the state-of-the-art in perceptual research used to build this model in detail as we outline our method in the following sections.

### 2.1 Quality metrics

Visual quality metrics are a mature field of research, and an important tool in research and development. They both help avoid costly user studies and lend to automation through optimization or as elements in cost functions for neural networks. The metrics that are the most relevant for our considerations are listed in Table 1. We will center our discussion around their principal approach and capabilities, listed in the columns of the table.

*Principal approaches to quality prediction.* A successful quality metric should predict the visual impact of distortions of different character and type. For example, if a denoising method needs to find a compromise between noise and blur, an image quality metric should indicate such a compromise, which is consistent with human judgment. Simple *signal quality* metrics, such as Peak Signal-to-Noise Ratio (PSNR), are known to fail in this task [Huynh-Thu and Ghanbari 2008]. Metrics based on different measures of *correlation* between the reference and distorted content (e.g. the Structural Similarity Index Measure or SSIM [2001]) are much more successful. Another statistical measure that has proven to correlate well with human quality judgment is the *entropy* of the distributions extracted from image content [Soundararajan and Bovik 2012]. A more fundamental, bottom-up approach involves modeling low-level vision based on *psychophysical models*, such as contrast sensitivity function (CSF). The benefit of relying on psychophysical models is that a

metric can account for physical properties of the display (brightness, size, viewing distance). Since the focus of our work is foveation with respect to a display, we follow that approach for our metric. Finally, the convolutional neural networks used in *machine learning* have been shown to serve as remarkably robust quality predictors when the activations values of their inner layers are used as features for comparing pairs of images [Zhang et al. 2018].

*Image metrics.* Canonical methods for assessing visual quality work by comparing a pair of images. Some of these metrics only consider per-pixel numerical differences, e.g. L1, L2, and PSNR, and do not account for spatial or temporal visual perception. Other metrics like SSIM [Wang et al. 2001] build on spatial image statistics and are often psychophysically calibrated [Mantiuk et al. 2011]. A recent class of image metrics utilizes deep image features in layers of deep neural networks trained for image-based tasks [Zhang et al. 2018]. While image-only metrics are popular for measuring image quality, they do not extend to videos. Specifically, although we can apply them to each frame of a video, the lack of inter-frame information makes these metrics unsuitable in identifying temporal visual artifacts. We demonstrate this in the accompanying video.

*Video metrics.* Assessing video quality requires different methods and assessing image quality, mostly due to interaction of spatial and temporal vision. For example, high frequency noise could be well visible in an image but it can disappear in high frame-rate video due to temporal integration of visual system. The notable examples of metrics that address temporal aspects are MOVIE [Seshadrinathan and Bovik 2009], STRRED [Soundararajan and Bovik 2012] and HDR-VQM [Narwaria et al. 2015]. MOVIE metric assesses spatial artifacts separately from motion artifacts. It decomposes test and reference video sequences using spatio-temporal Gabor filter bank and estimates optical flow on the reference video to analyze differences along motion directions. Motion artifacts are assumed to cause deviations from reference motion, which are found by analyzing the Gabor channel amplitude responses in the frequency domain. STRRED is Spatio-Temporal Reduced Reference Entropy Difference, which estimates the quality degradation by calculating the entropy difference between reference and distorted video sequences. The entropy is computed from the distribution of wavelet coefficients, which is modeled as Gaussian Scale Mixture. The entropy difference is evaluated in non overlapping blocks, separately for spatial and temporal dimensions. STRRED was one of the few metrics that could predict some aspects of temporal quality in our new dataset. We also included in our analysis HDR-VQM, which is one of the most popular metrics for HDR video that accounts for physical calibration. The metric decomposes video into spatial bands, which are then split into spatio-temporal "tubes". The distortion is estimated by two stage temporal pooling, which uses the percentile of the largest distortions to focus on most salient artifacts. None of those video metrics models the visibility of high temporal frequency artifacts associated with flicker, which are critical for assessing the quality of foveated rendering methods.

*Foveated metrics.* The lower sensitivity to artifacts appearing outside the fovea has been modeled in a number of foveated metrics.

<sup>1</sup>Source code at: <https://github.com/gfxdisp/FovVideoVDP>

Table 1. Quality metrics and their capabilities. Columns indicate whether a metric has been designed for video, foveated viewing, operates on photometric units (accounts for display brightness) and physical size of the display. We also mention the main approach of each metric and the datasets that were used to calibrate/validate the metric in the original paper.

Metrics	Video	Foveated	Photometric	Disp. geometry	Approach	Calibration dataset
PSNR	No	No	No	No	Signal quality	N/A
MS-SSIM [Wang et al. 2003]	No	No	No	No	Correlation	LIVE
HDR-VDP-3 [Mantiuk et al. 2011]	No	No	Yes	Yes	Psychophysical model	UPIQ
FA-MSE/SSIM [Rimac-Drlje et al. 2011]	No	Yes	No	Yes	Correlation	LIVE video
FWQI [Wang et al. 2001]	No	Yes	No	Yes	Correlation	None
HDR-VDP2-FOV [Swafford et al. 2016]	No	Yes	Yes	Yes	Psychophysical model	Own (3 stimuli)
Contrast-Aware [Tursun et al. 2019]	No	Yes	Yes	Yes	Psychophysical model	Own experiment
HDR-VQM [Narwaria et al. 2015]	Yes	No	Yes	Yes	Psychophysical + corr.	Own dataset (10 videos)
MOVIE [Seshadrinathan and Bovik 2009]	Yes	No	No	No	Gabor filterbank	VQEG FRTV Phase 1
STRRED [Soundararajan and Bovik 2012]	Yes	No	No	No	Entropy difference	LIVE video
LPIS [Zhang et al. 2018]	No	No	No	No	Machine learning	BAPPS (own)
<b>FovVideoVDP (ours)</b>	Yes	Yes	Yes	Yes	Psychophysical model	FovDots (own) + UPIQ + Deep-Fovea + LIVE-FBT-FCVR

FWQI (Foveated Wavelet Quality Index) is a correlation-based index in which image locations at large eccentricities are assigned lower sensitivity. FA-MSE and FA-SSIM metrics [Rimac-Drlje et al. 2011, 2010] extend the popular MSE and SSIM metrics by weighting them by a sensitivity function. The sensitivity function combines the eccentricity term from [Wang et al. 2001] with a new term that decreases at high retinal velocities. Swafford et al. [2016] extended HDR-VDP-2 to account for eccentricity by scaling the contrast sensitivity by the cortical magnification factor [Virsu and Rovamo 1979]. More recently, Tursun et al. [2019] proposed a metric specifically targeting foveated rendering, allowing sampling rate to vary with both eccentricity and spatial image content. The effect of eccentricity is modeled using the contrast sensitivity function proposed by Peli et al. [1991], which had to be modified to fit the new data. In contrast with the two latter works, we predict the effect of eccentricity while relying on existing contrast sensitivity functions [Kelly 1979; Laird et al. 2006; Mantiuk et al. 2020] and models of cortical magnification [Dougherty et al. 2003; Virsu and Rovamo 1979].

*Physically calibrated metrics.* It is important to make a distinction between metrics that operate on stimuli specified in physical units and those that operate on gamma-encoded images. Most popular image quality metrics, such as PSNR and MS-SSIM [Wang et al. 2003], take as input a pair of gamma-encoded images and ignore all aspects of their physical presentation, including screen resolution, size, viewing distance, and display peak luminance. This simplification makes these metrics easier to use but ignores important factors that affect image quality. Display-independent quality measures are no longer sufficient in an era where we routinely encounter displays with diverse characteristic. Physically-calibrated metrics, such as VDP [Daly 1993] or HDR-VDP [Mantiuk et al. 2011], require full physical specification of the input images, including effective resolution in pixels per visual degree and images calibrated in absolute units of luminance ( $\text{cd}/\text{m}^2$ ). Our proposed metric belongs to this latter category, as we aim to account for displays of varying sizes and peak-luminance levels. Finally, note that some metrics, such as FWQI, account for the display’s spatial specification (size, resolution and viewing distance) but do not model its photometric properties.

As we focus on methods that could be used as optimization criteria, we do not analyze the metrics intended to produce visual difference maps [Aydin et al. 2010; Daly 1993; Wolski et al. 2018; Ye et al. 2019] but that do not offer single-valued quality predictions.

## 2.2 Applications

A foveated spatio-temporal metric has several applications in visual computing, which we will discuss here.

*Foveated compression and rendering.* Mobile head-mounted displays are often limited in the quality and complexity of their content due to the technical challenges of generating or transmitting large field-of-view videos. To overcome this constraint, foveated video compression [Geisler and Perry 1998] has become an active field of research, and could potentially bring benefits beyond VR/AR. Recent methods like Deep Fovea [2019] reduce the peripheral resolution of an image and train a semi-supervised adversarial network to reconstruct missing details, reducing the number of pixels required to render or transfer, while striving to maintain spatio-temporal consistency.

Another crucial scenario for immersive displays is foveated rendering, where foveation is used in real-time to significantly reduce the computational cost. Guenter et al. [2012] introduced the first real-time method, using gaze tracking to display and degrade the rendered image resolution for concentric rings around the gaze location. Three layers are rendered with different pixel densities, then combined with a soft stepping function to avoid a sharp step at the boundaries. Patney et al. [2016] use a gaze-tracker-equipped VR headset and achieve foveation through variable shading, where visibility is computed at full resolution across the whole visual field, but materials and lighting are evaluated at lower rates in the periphery. This reduces the pixel shading cost by up to 70%. While foveated rendering has come a long way in the past decade, there is no proven visual metric to evaluate new algorithms, or to compare existing solutions for novel AR/VR hardware.

*Other Applications.* A well-calibrated video metric is also useful in predicting spatio-temporal visual differences between videos in a general, non-foveated scenario. For example, it can be used to

measure the visibility of temporal motion artifacts due to limited frame rate, which we demonstrate in Section 6.2. Further, in predicting visual differences for varying fields-of-view, such a metric can account for viewing conditions like display size and viewing distance, providing a more accurate estimate of whether and which artifacts would be most visible to the end users. We demonstrate this in Section 6.1, where we show how our proposed metric can account for viewing distance.

Subtle Gaze Direction [Bailey et al. 2009] is another application for a foveated video metric. This method uses subtle peripheral modulations of an image or video to direct a user’s attention, and has utility in fields like medical image analysis [Sridharan et al. 2012] and VR redirected walking [Sun et al. 2018]. Since the method relies on spatio-temporal peripheral feedback, a foveated video metric is uniquely applicable. Specifically, as shown in Section 6.4, it can help calibrate peripheral image manipulations so that they are sufficiently visible to capture a user’s attention, but not too distracting.

### 3 FOVEATED VIDEO QUALITY METRIC

Our goal is to design a full-reference visual difference metric, which models the major stages of the early visual system. We aim to make this metric as simple as possible while modeling the relevant aspects of low-level vision. Finally, it is valuable for the resulting metric to be differentiable, so it may be implemented as a loss function for reconstruction algorithms. Unlike some other metrics modeling low-level vision, such as the Visual Difference Predictor (VDP) [Daly 1993], which model just-detectable difference (detection and discrimination), our focus is on quantifying supra-threshold differences. This is an important requirement, as threshold metrics do not produce discriminative results when comparing inputs that contain significant differences, and so can be unsuitable for use as an optimization criterion. Furthermore, the focus of this work is on obtaining a single-value quality score that would correlate well with psychophysical measurements and guide optimization of foveated rendering methods according to a perceptual criterion. This single quality value is scaled in interpretable units of (JOD, explained in Section 3.9), which corresponds to the increase in preference across the population.

*Limitations.* To ensure low complexity, we did not consider several processing stages that can be found in more complex metrics. We do not model glare due to scattering of light in the optics of the eye and on the retina as it involves convolutions with large kernels. We also do not model orientation-selective visual channels and cross-channel masking as that would further increase both processing and memory overhead. We do not model the loss of sensitivity due to eye movements as most datasets do not provide this information. Color vision is currently not modeled by FovVideoVDP because there is only scarce data available for color contrast sensitivity across eccentricities and temporal frequencies. However, chromatic distortions are typically much less noticeable than luminance distortions [Winkler et al. 2001].

*Overview.* A schematic of our method is shown in Figure 2. Our metric operates on inputs described in physical units of luminance (videos or images), which are obtained by employing a display model.

In addition, the distance from the display is used to model the viewers’ effective retinal resolution. The luminance map is decomposed into sustained and transient temporal channels, which are then converted to Laplacian pyramids. Next, each spatial-frequency pyramid band is encoded in units of physical contrast, and then passed to the masking model, which estimates the perceived difference for each band. These values can be used to produce a per-pixel visual difference map or pooled to obtain a single-valued quality score, which is then scaled in just-objectionable-difference (JOD) units.

The following sections follow the flow of Figure 2, explaining each step in detail and contextualizing with respect to the state-of-the-art. While we investigated many variations of this metric, the following sections describe the version with the best performance. Other variants are discussed in Section 5.4. We currently do not model display blur (MTF) and VR lens aberrations, which could be relevant for some displays.

#### 3.1 Display model

Since the visual models we employ are specified in physical units, we need to linearize the input pixel values so they are represented in units of luminance. Inputs can be provided in any color space, such as ITU-R Rec. BT.709 RGB with gamma-encoding, or ITU-R Rec. BT.2100 RGB with PQ coding. We convert input values by employing a display model, consisting of its basic characteristics (peak luminance, contrast and color space), to transform the input pixel values into the luminance emitted from a given display, scaled in absolute units of  $\text{cd}/\text{m}^2$ . When inputs are given in standard dynamic range (SDR), we use a gain-gamma-offset (GOG) display model [Berns 1996].

*Angular display resolution.* To accurately model spatial effects, we need to convert distances and frequencies specified in display space into units on the retina. Because of the approximately spherical shape of the eye, we express the distances in visual degrees and frequencies in cycles per visual degree. For displays spanning a small field-of-view, the angular resolution in pixels per visual degree can be approximated as:

$$n_{\text{ppd},0} = \frac{\pi}{360 \operatorname{atan}\left(\frac{0.5 d_{\text{width}}}{r_h d_v}\right)}, \quad (1)$$

where  $d_{\text{width}}$  is the display width in meters,  $r_h$  is the horizontal resolution in pixels and  $d_v$  is the viewing distance in meters. However, since our focus is on wide-field-of-view VR/AR displays, we need to account for the changes in angular resolution with the eccentricity (the viewing angle relative to central view direction). The angular resolution changes with eccentricity  $e$  (in visual degrees) as:

$$n_{\text{ppd}}(e) = n_{\text{ppd},0} \frac{\tan\left(\frac{\pi e}{180} + 0.5 n_{\text{ppd},0}^{-1}\right) - \tan\left(\frac{\pi e}{180}\right)}{\tan\left(0.5 n_{\text{ppd},0}^{-1}\right)}. \quad (2)$$

Figure 3 shows how the angular resolution changes with the eccentricity for different displays. The change is particularly substantial for VR/AR displays, which tend to have low angular resolution and wide field of view.

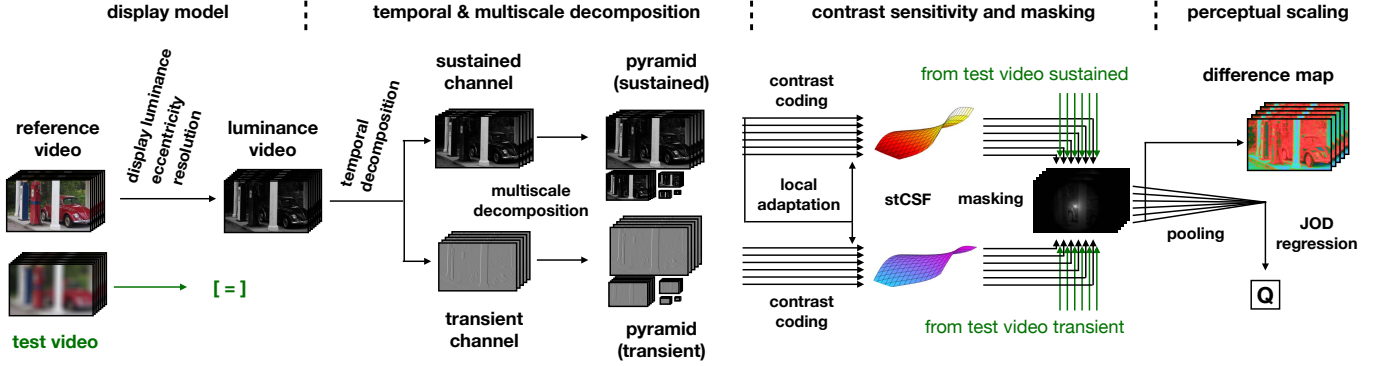


Fig. 2. This figure shows a diagram of the method proposed in this work. The test and reference videos are processed in the same manner up to the *masking* model block, where the perceived difference between the two is evaluated.

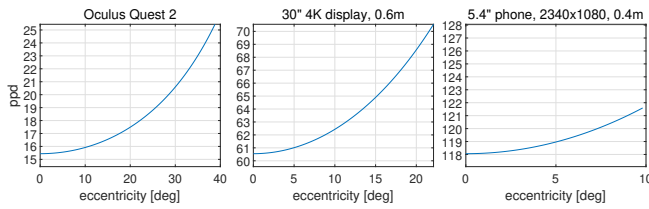


Fig. 3. The retinal resolution (in pixels-per-degree) of several displays as a function of eccentricity. There is a substantial change in resolution for displays that span a large field-of-view (e.g. Quest 2 and a 4K display), which needs to be accounted for in the visual model.

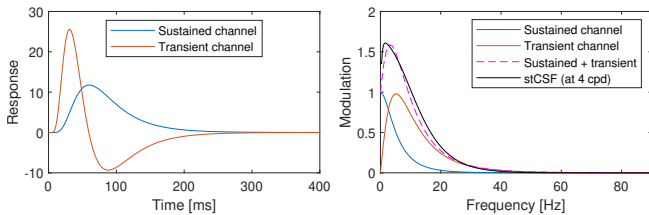


Fig. 4. Left: impulse response functions for the two temporal channels. Right: The same impulse response functions are plotted in the frequency domain. The sum of both channels is close to the spatio-temporal CSF [Laird et al. 2006], shown as a black line.

### 3.2 Temporal channels

It has been argued that the visual system encodes temporal changes in a pair of temporal channels [Burbeck and Kelly 1980; Hammett and Smith 1992]: a *sustained channel* encodes slow changes, and a *transient channel* encodes fast changes. Smith [1998] models the response of the sustained system as a cascade of exponential filters. We observe that the same characteristic can be obtained with a simpler Gaussian function of logarithmic time  $t$  (in seconds):

$$R_S(t) = k_1 \exp\left(\frac{(\log(t + \epsilon) - \log(\beta_S))^2}{2\sigma_S^2}\right) \quad (3)$$

where  $\beta_S = 0.06$  [s] represents the offset (lag) of the response and  $\sigma_S = 0.5$  controls the bandwidth of the filter. The parameters were

selected to match the functions from [Smith 1998]. The normalization constant  $k_1 = 0.00573$  ensures that the filter response to a static (0 Hz) signal is not affected and is equal to the inverse of the integral of the exponential function.  $\epsilon = 0.0001$  is a small constant that ensures the logarithm is finite at  $t = 0$ . As in [Smith 1998], we model the response of the transient channel as the first derivative of the response of the sustained channel:

$$R_T(t) = k_2 \frac{R_S}{dt}(t) = k_2 \frac{-R_S(t) (\log(t + \epsilon) - \log(\beta_S))}{\sigma_S^2 (t + \epsilon)} \quad (4)$$

The impulse response functions for both channels and their Fourier transforms are plotted in Figure 4. The plot shows that the peak frequency response for sustained is at  $f_S = 0$  Hz and at  $f_T = 5$  Hz for the transient channel. We will later use those values to model sensitivity. The normalization constant  $k_2 = 0.0621$  ensures that the contrast at the peak frequency of 5 Hz is preserved.

The dashed line on the right side of Figure 4 shows a 4 cpd slice of the spatio-temporal contrast sensitivity function (stCSF) [Laird et al. 2006] plotted as a dotted black line. It can be seen that the sum of both sustained and transient channels matches the stCSF well, confirming our choice of filters. We generate digital filters broad enough to cover the non-zero portions of the temporal filters (250 ms), which we then apply as sliding windows.

### 3.3 Multi-scale decomposition

Both psychophysical data [Foley 1994; Stromeyer and Julesz 1972] and neuropsychological recordings [De Valois et al. 1982] show evidence for the existence of mechanisms that are selective to narrow bands of spatial frequencies and orientations. To mimic the decomposition that happens in the visual cortex, visual models commonly employ multi-scale image decompositions, such as wavelets or pyramids [Simoncelli and Freeman 2002], or band-pass filters in the Fourier domain [Daly 1993; Watson 1987]. Considering that such a decomposition is one of the most computationally expensive parts of a visual model, we employ the decimated Laplacian pyramid [Burt and Adelson 1983], which can be efficiently computed and stored. The main drawback of the Laplacian pyramid is that it does not isolate patterns of different orientation, which are known to contribute differently to visual masking [Foley 1994] and sensitivity

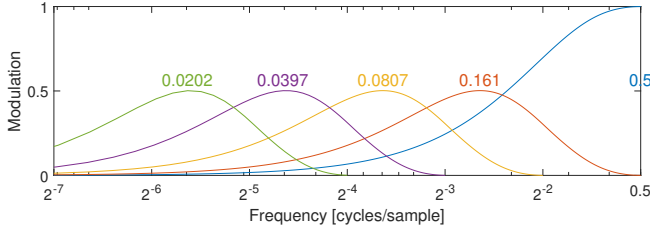


Fig. 5. Spatial frequency bands of the Laplacian pyramid [Burt and Adelson 1983]. The colored numbers indicate the frequency at the peak of each band.

[Barten 2004]. However, we found that such orientation-selectivity has little impact on the predictions made for complex images.

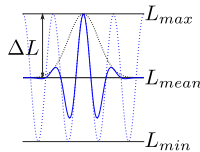
Figure 5 shows the frequency response of the filters and the consecutive levels of the Laplacian pyramid found by the discrete Fourier transform of the corresponding filters. The numbers above each band indicate the peak frequency of each band. It is worth noting that the highest frequency band has twice the amplitude of that of the following bands. This is because all other bands are computed as a difference of two low-pass filtered bands. Another important observation is that the peak frequencies cannot be obtained with halving by the Nyquist frequency of 0.5 samples/cycle, as commonly assumed. All those nuances must be accounted for to ensure that the response is scaled in the correct units. We model the peak frequency of each band in cycles-per-degree (cpd) as:

$$\rho_b = \begin{cases} 0.5 n_{ppd} & \text{if } b = 1 \\ \frac{0.1614}{2^{b-2}} n_{ppd} & \text{otherwise} \end{cases}, \quad (5)$$

where  $n_{ppd}$  is the angular image resolution given in pixels per visual degree. We select the height of the pyramid so that the lowest frequency is at least 0.5 cpd. This is because lower frequencies are not relevant for the types of distortions we consider and also because distortions in low frequencies cannot be well localized, leading to issues with pooling across multiple bands.

### 3.4 Contrast coding and local adaptation

Psychophysical models of contrast sensitivity, which we will discuss in the next section, are typically defined in terms of Michelson contrast and expressed as:

$$C = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} = \frac{\Delta L}{L_{mean}}. \quad (6)$$


For sine gratings and Gabor patches, such as those shown on the right, Michelson contrast is equal to Weber contrast on the right side of the equation. Our goal is to find a local measure of such contrast in complex images. Furthermore, for efficiency, we want this measure to be easily computed from the Laplacian and Gaussian pyramids, which we use for multiscale decomposition.

The coefficients of the Laplacian pyramid encode information about the amplitudes of the band-pass filtered signal and, therefore, approximate  $\Delta L$  from Eq. 6. Kingdom and Whittle [1996] argued that that contrast discrimination thresholds can be well explained

by the Weber law ( $\Delta C/C \approx \text{const.}$ ) when the denominator in Eq. 6 represents the adapting luminance rather than the mean luminance. The adapting luminance is mostly influenced by a small neighborhood of a given image location of around 0.5 deg [Vangorp et al. 2015]. For efficiency reasons, we approximate this region using the value from the Gaussian pyramid at the level that is one higher than the given level of the Laplacian pyramid. We can then express the local contrast as:

$$C_{b,c}(\mathbf{x}) = \frac{\mathcal{L}_{b,c}(\mathbf{x})}{\mathcal{G}_{b+1,S}(\mathbf{x})} = \frac{\mathcal{L}_{b,c}(\mathbf{x})}{L_a(\mathbf{x})} \quad (7)$$

where  $\mathcal{L}_{b,c}(\mathbf{x})$  is the coefficient of the Laplacian pyramid at the pixel coordinates ( $\mathbf{x}$ ), pyramid level  $b$  and temporal channel  $c$  (sustained or transient).  $\mathcal{G}_{b+1,S}(\mathbf{x})$  is the corresponding coefficient of the Gaussian pyramid for the sustained channel. The contrast encoding above is similar to the local band-limited contrast proposed by [Peli 1990], except that, for efficiency, we rely on Laplacian and Gaussian pyramids rather than Fourier-domain cosine log filters, we define contrast for temporal channels and we use a higher level of the Gaussian pyramid to account for local adaptation.

It should be noted that we do not model photoreceptor non-linearity (luminance masking) as done in many other visual models [Daly 1993; Mantiuk et al. 2011]. We attempted introducing this non-linearity but we found a degradation in the performance as compared with simple contrast encoding. The effect of luminance is modeled in the contrast sensitivity function, as explained in the next section.

### 3.5 Spatio-temporal contrast sensitivity

The CSF is a psychophysical model that describes the smallest contrast,  $C$ , that is detectable by an average observer. The CSF predicts the sensitivity  $S$ , which is defined as the inverse of the contrast detection threshold  $C_T$ :

$$S(\rho, \omega, L_a, a) = \frac{1}{C_T} = \frac{L_a}{\Delta L_T}, \quad (8)$$

where  $\Delta L_T$  is the smallest detectable luminance difference and  $L_a$  is the background/adaptation luminance. In our metric, we need to predict sensitivity as a function of the spatial frequency,  $\rho$ , in cycles per degree, the temporal frequency,  $\omega$ , in Hz, the adapting/background luminance,  $L_a$ , in  $\text{cd/m}^2$ , and the size,  $a$ , in  $\text{deg}^2$ . The size parameter is particularly important for modeling extrafoveal vision, which we will discuss in the next section. The sensitivity is also affected by the orientation of the spatial stimulus (horizontal, vertical and diagonal), which we ignore because it is a much weaker factor and because orientations are not isolated in our multiscale decomposition.

Unfortunately, there is no existing model that accounts for all the required dimensions. Popular CSF models, such as that of Barten [1999] or Daly [1993], do not account for temporal frequency. Kelly's spatio-temporal CSF [Kelly 1979; Laird et al. 2006] accounts for spatial and temporal frequencies, but does not account for luminance and the size of the stimulus. To build a model that accounts for all the factors, we rely on approximate independence between spatial/temporal frequencies and other dimensions [Watson and Ahumada 2016], and combine two sensitivity models: the Kelly-Daly spatio-velocity CSF [Laird et al. 2006] ( $S_{sv}$ ) with a recent model of

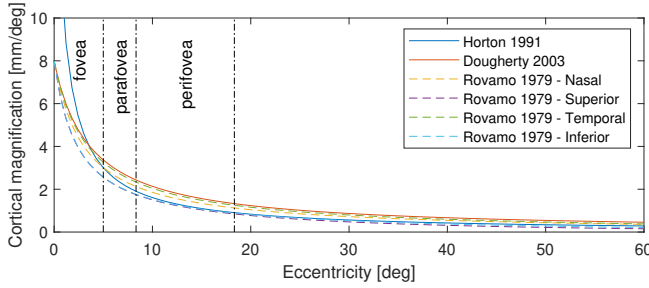


Fig. 6. The magnitude of cortical magnification according to several models [Dougherty et al. 2003; Horton 1991; Rovamo and Virsu 1979; Virsu and Rovamo 1979]. Rovamo et al. model provides different estimates for each part of the visual field.

spatio-chromatic contrast sensitivity modeled up to 10,000 cd/m<sup>2</sup> [Mantiuk et al. 2020] ( $S_{sc}$ ):

$$S_{fov}(\rho, \omega, L_a, a) = S_{sc}(\rho, L_a, a) \frac{S_{sv}(\rho, \omega/\rho)}{S_{sv}(\rho, 0)} = S_{sc}(\rho, L_a, a) \frac{S_{st}(\rho, \omega)}{S_{st}(\rho, 0)}. \quad (9)$$

As noted by Daly [1998], the spatio-velocity CSF ( $S_{sv}$ ) is equivalent to the spatio-temporal CSF ( $S_{st}$ ) when velocity is set to  $v = \omega/\rho$ . The combined CSF relies on the Kelly-Daly CSF to model the relative change of sensitivity due to temporal frequency and predicts the effect of other factors using the spatio-chromatic CSF. The obtained CSF predicts the sensitivity for foveal vision when the central portion of the retina is used to perceive a stimulus. In the next section, we will extend the model to account for extrafoveal vision.

### 3.6 Cortical magnification and peripheral sensitivity

To model foveated vision, we need to understand how the sensitivity changes outside the foveal part of the visual field. It has been argued that the changes in detection threshold can be explained by the concept of *cortical magnification*. This model describes how many neurons in the visual cortex are responsible for processing a particular portion of the visual field [Horton 1991]. The central, foveal region is processed by many more neurons (per steradian of visual field) than the extrafoveal region. The cortical magnification is expressed in millimeters of cortical surface per degree of visual angle. Factors of cortical magnification for several models proposed in the literature are plotted in Figure 6. We rely on the model by Dougherty et al. [2003], which was fitted to fMRI measurements of V1. The cortical magnification is modeled as:

$$M(e) = \frac{a_0}{e + e_2}, \quad (10)$$

where  $e$  is eccentricity in visual degrees and the fitted parameters are  $a_0 = 29.2$  mm and  $e_2 = 3.67^\circ$ . Virsu and Rovamo [1979; 1979] showed that the differences in detection of sinusoidal patterns and also discrimination of their orientation or direction of movement, can be compensated by increasing the size of the stimuli in the peripheral vision and the size increase is consistent with the inverse of cortical magnification. We follow that observation to extend our CSF model (Eq. 9) to extra-foveal spatio-temporal CSF, with the extra parameter of eccentricity. The extended extra-foveal model modulates both the frequency and size of the stimulus using the

relative cortical magnification factor  $M_{rel}$ :

$$S_{exfov}(\rho, \omega, L_a, e) = S_{corr} \cdot S_{fov} \left( \frac{\rho}{M_{rel}(e)}, \omega, L_a, \pi (\sigma_s M_{rel}(e))^2 \right). \quad (11)$$

where  $M_{rel}(e) = (M(e)/M(0))^{k_{cm}}$  and  $k_{cm}$  is a free parameter. We model the size of the stimulus as the area of a disk that is modulated by the inverse of the cortical magnification (the last parameter of  $S_{fov}$ ). It should be noted that as the size of the stimulus increases, its spatial frequency decreases proportionally, therefore the spatial frequency  $\rho$  is modulated by the relative cortical magnification  $M_{rel}(e)$ . Furthermore, since lower frequencies are detected by larger receptive fields, the size of the stimulus,  $\sigma_s$ , varies with frequency:

$$\sigma_s = \frac{\sigma_0}{\rho}, \quad (12)$$

where  $\sigma_0$  is a free parameter. Here we model the area of the stimulus as the area of a disk with the radius given by  $\sigma_s M_{rel}(e)$ . Finally, we also add a sensitivity correction factor  $S_{corr}$ , which we use as a free parameter to adjust the sensitivity of our metric. Our CSF for several eccentricities is compared Daly's CSF in Figure 7 (top). While the trends are similar, the shape of our function at lower frequencies is closer to the data reported by Virsu and Rovamo [1979, fig. 2c]. In Figure 7 (bottom) we show our CSF as the function of luminance and eccentricity for both sustained and transient temporal channels. These functions show how the sensitivity can vary inside each spatial-frequency band as both the adapting luminance and eccentricity can be different for each pixel position. Our CSF is a 4-dimensional function, however, as we rely on only two temporal frequencies (0 Hz and 5 Hz), it can be discretized as two 3D look-at-tables, or a family of 2D look-up tables, with an individual 2D table for each spatial and temporal band. If no foveated viewing is desired and the metric should operate as other non-foveated metrics (HDR-VDP, MS-SSIM, etc.), we set the eccentricity  $e = 0$ .

### 3.7 Contrast masking

Contrast is less visible when superimposed on another contrast due to the phenomenon known as contrast masking [Foley 1994; Legge and Foley 1980; Watson and Solomon 1997]. This effect is illustrated in Figure 8, in which a Gabor patch is more difficult to detect when presented on the background of the same spatial frequency. The model of contrast masking in our metric is responsible for transforming a pair of band-limited contrast values, coming from test and reference images, into perceived differences. It is the core component of the metric and much of the metric performance depends on the execution of that processing block. We have experimented with several masking models, some of them compared in Section 5.4, but here we describe the model that resulted in the best performance.

The majority of the masking models found in the literature operate on contrast that has been normalized by the detection threshold:

$$C'_{b,c}(\mathbf{x}) = C_{b,c}(\mathbf{x}) S_{b,c}(\mathbf{x}) \quad (13)$$

where the contrast  $C_{b,c}$  is given by Eq. (7). Daly [1993] has shown that such a normalization by the sensitivity helps to eliminate the variations in the shape of the masking function between spatial frequencies. The contrast sensitivity is given by:

$$S_{b,c}(\mathbf{x}) = S_{exfov}(\rho_b(\mathbf{x}), \omega_c, L_a(\mathbf{x}), e(\mathbf{x})), \quad (14)$$

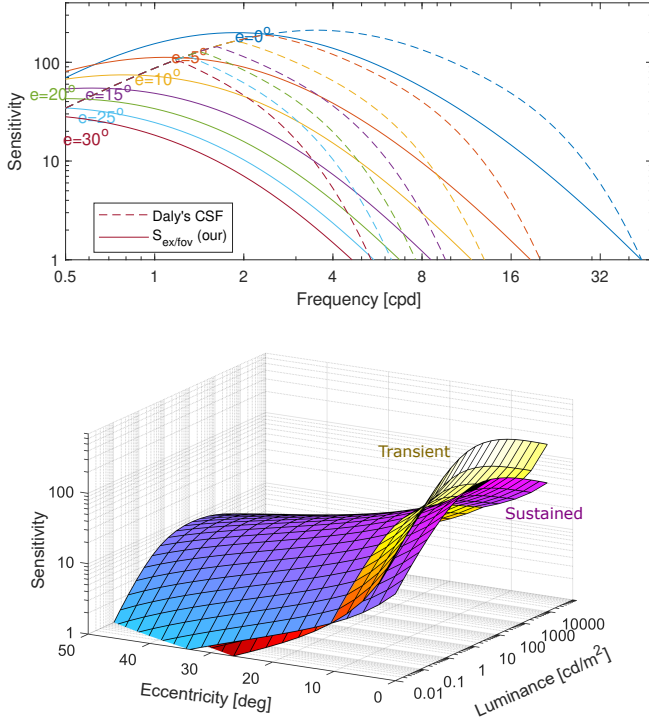


Fig. 7. *Top*: The contrast sensitivity as a function of frequency and eccentricity. Our function, based on the cortical magnification factor, is compared with that of Daly [1993]. *Bottom*: The same contrast sensitivity as the function of luminance and eccentricity for two temporal channels. The function is plotted for the band with the peak frequency of 1 cpd. The transient channel is more sensitive in that band at low to medium eccentricities and the sustained channel is more sensitive at higher eccentricities.

where  $\rho_b(\mathbf{x})$  is the peak spatial frequency of the band  $b$  (Eq. (5)) at the coordinates  $(\mathbf{x})$  (spatial frequency varies with position due to screen projection — Eq. (2)),  $\omega_c$  is the peak temporal frequency of the sustained or transient channel (0 Hz or 5 Hz, see Section 3.2),  $L_a(\mathbf{x})$  is the local luminance of adaptation (see Section 3.4) and  $e(\mathbf{x})$  is the eccentricity at the pixel coordinates  $\mathbf{x}$ . It is worth noting that multiplication by the sensitivity expresses the contrast as a multiple of threshold contrast:  $C \cdot S = \Delta L / L_a \cdot L_a / \Delta L_T = \Delta L / \Delta L_T$ .

From several variants of the masking models we tested, the best performance was achieved by the model that encoded the perceived difference between a pair of test ( $C'_{b,c}{}^{\text{test}}$ ) and reference ( $C'_{b,c}{}^{\text{ref}}$ ) contrast values as:

$$D_{b,c}(\mathbf{x}) = \frac{\left| C'_{b,c}{}^{\text{test}}(\mathbf{x}) - C'_{b,c}{}^{\text{ref}}(\mathbf{x}) \right|^p}{1 + (k C_{b,c}^{\text{mask}}(\mathbf{x}))^{q_c}} \quad (15)$$

where  $p$ ,  $q_c$  and  $k$  are the parameters of the model. The masking parameter  $q_c$  was found separately for the sustained and transient channels ( $q_S$  and  $q_T$ ). The mutual masking signal (see [Daly 1993, p.192]) is given by:

$$C_{b,c}^{\text{mask}}(\mathbf{x}) = \min \left\{ \left| C'_{b,c}{}^{\text{test}}(\mathbf{x}) \right|, \left| C'_{b,c}{}^{\text{ref}}(\mathbf{x}) \right| \right\}. \quad (16)$$

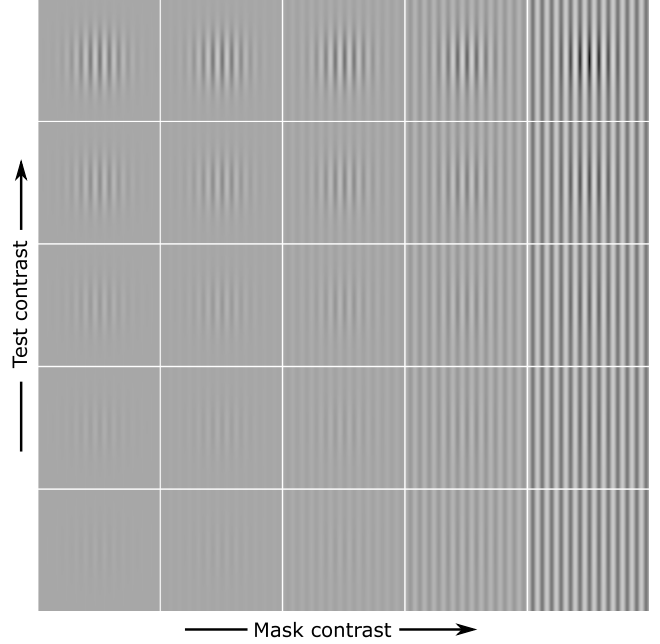


Fig. 8. Illustration of the contrast masking effect. The contrast of a masker (sine grating) is increasing from left to right, making the test contrast (Gabor) more difficult to detect.

We experimented with variants that included local summation of the masking signal (convolution with a Gaussian), but we did not achieve a noticeable improvement in performance.

### 3.8 Pooling

Once we have computed the perceived difference measures, we need to pool the values across all coefficients in each band, across spatial frequency bands ( $b$ ), across temporal channels ( $c$ ) and finally across all the frames ( $f$ ):

$$D_{\text{pooled}} = \frac{1}{F^{1/\beta_f}} \left\| w_c \left\| \frac{1}{N_b^{1/\beta_x}} \| D_{b,c}(\mathbf{x}) \|_{\beta_x, \mathbf{x}} \right\|_{\beta_b, b} \right\|_{\beta_c, c} \right\|_{\beta_f, f}, \quad (17)$$

where  $\|\cdot\|_{p,v}$  is a  $p$ -norm over the variable  $v$ :

$$\|f(v)\|_{p,v} = \left( \sum_v |f(v)|^p \right)^{1/p}. \quad (18)$$

$w_c$  is the weight of sustained or transient component. It should be noted that the  $p$ -norms across the coefficients and frames are normalized by the number of coefficients in each band ( $N_b$ ) and the number of frames ( $F$ ). This is because we do not want the visual error to grow with the resolution and the number of frames. The exponents in the  $p$ -norms ( $\beta_f$ ,  $\beta_c$ ,  $\beta_b$  and  $\beta_x$ ) are related to the slope of the psychometric function, assuming that the pooling represents probability summation [Robson and Graham 1981]. All those coefficients are optimized parameters in our model.



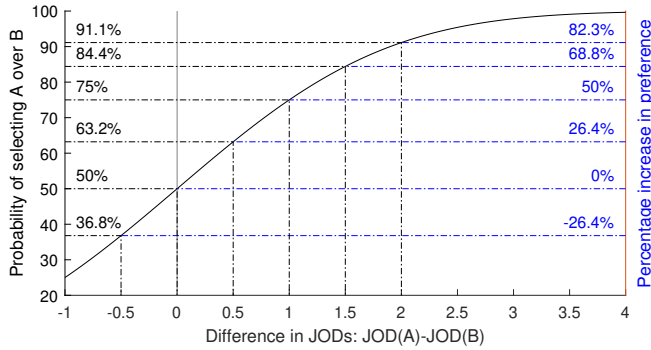


Fig. 9. Mapping of difference in quality between condition A and B in JOD units to interpretable quality difference. The difference of 1 JOD means that 75% of the population will select condition A over B. This corresponds to a relative increase in preference of 50% (relative to a random choice:  $p_{inc} = \frac{p(A>B)}{0.5} - 1$ ).

### 3.9 JOD regression

In the final step, we regress the pooled visual difference into just-objectionable-difference (JOD) units:

$$Q_{JOD} = 10 - \alpha_{JOD} (D_{pooled})^{\beta_{JOD}}, \quad (19)$$

to obtain the final quality of the video sequence.  $\alpha_{JOD}$  and  $\beta_{JOD}$  are optimized parameters. Following the tradition of quality indices, the JOD units increase with quality. The highest quality, reported for no difference between a pair of content, is anchored at 10 JODs to avoid negative values.

The difference between JOD and more commonly known just-noticeable-difference (JND) units is that the former represents the difference with respect to the reference image while the latter represents the difference between a pair of images [Perez-Ortiz et al. 2020, Fig.5]. Two distorted images could be very different from each other in terms of JNDs but they can have similar JODs with respect to the reference. The main advantage of the JOD units is that they provide an interpretable scale of quality for which we can estimate the increase in preference across the population. For example, if method A resulted in image difference of 9.5 JOD and method B in 8.5 JOD, we can interpret this difference of 1 JOD as 50% increase in preference for method A (over a random choice). This is illustrated for other differences in JODs in Figure 9.

*Difference maps.* In some application it is desirable to know not only the overall level of distortion as an JOD value, but also how the distortions are distributed within an image or video. For that purpose, we extract the visual difference map by reconstructing each frame ( $D_{rec}$ ) from the Laplacian pyramid coefficients  $D_{b,c}$ . Then, we use the same JOD regression coefficients as in Eq. (19) but we represent increasing distortions levels rather than quality:

$$D_{map}(x) = \alpha_{JOD} (D_{rec}(x))^{\beta_{JOD}}. \quad (20)$$

Unlike the visual difference maps produced by other metrics [Mantiuk et al. 2011; Wolski et al. 2018; Ye et al. 2019], which saturate when the distortions are well visible, our maps represent both near-threshold and supra-threshold distortions, scaled in JOD units. We

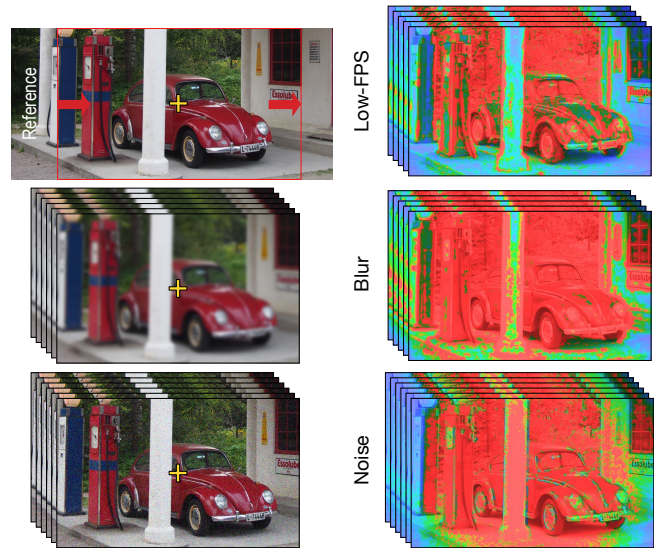


Fig. 10. This figure shows a sample output of our method for a simple video consisting of a sideways panning inside of a larger image (shown on the top left). The pan has a speed of 5 deg/s, and the video is sampled at 120 Hz. This reference is compared against the same video, but at a lower 30 Hz frame rate, with the difference map shown on the top right. In addition, we compare against a version of the 120 Hz video with added blur and temporal noise, shown on the 2nd and 3rd rows respectively. Note the fall-off in response strength away from the fixation point, set to the center of the image in all three rows and marked with a cross. As expected, in the top row differences are most visible around vertical edges. In the middle row, high frequencies everywhere are significantly distorted, while the opposite is true in the bottom row where the noise is most visible in flat areas.

use the color coding scheme with the contextual image from HDR-VDP-3. Some examples of such difference map can be found in Figures 1, 10 and 20.

### 3.10 Implementation details, timings

The metric has been initially implemented and tested in Matlab, then ported to PyTorch. The Matlab implementation used `gpuArrays` to move the processing to CUDA cores on the GPU. The temporal filter was implemented as a sliding window, so that only a fixed number of frames had to be present in the GPU memory. The most computationally expensive part of the metric was the construction of the Laplacian pyramid. We could accelerate this part by forming a  $width \times height \times 4$  image from sustained and transient channels of test and reference images and decomposing them all into the pyramid in a single step (instead of 4).

The execution times of both implementations of our metric, for images and video can be found in Table 2. The execution times of FovVideoVDP are substantially shorter than for other metrics of similar complexity (see the supplementary). The short execution times of FovVideoVDP make it practical for processing images of large resolution and video.

Table 2. Run time performance of our metric, measured for various input sizes and on both MATLAB and Pytorch. Measurements taken on a computer with Intel Core i7-7800X CPU and NVIDIA GeForce RTX 2080 GPU.

Resolution	Frames	Metric Time	
		MATLAB	Pytorch
1280×720	1	91.07 ms	65.11 ms
1920×1080	1	119.82 ms	93.61 ms
3840×2160	1	214.94 ms	242.13 ms
1280×720	60	3.71 sec	3.43 sec
1920×1080	60	5.17 sec	5.51 sec
3840×2160	60	23.61 sec	14.68 sec

#### 4 FOVEATED RENDERING DATASET (FOVDOTS)

Our goal is to obtain a metric suitable for foveated rendering. However, there are very few relevant datasets that could help us evaluate such a metric, and even these have severe limitations (see Section 5.1). Specifically, they either contain overly simple stimuli (such as sine gratings), which are not representative of the content used in the target applications of foveated video transmission and rendering. Other datasets contain a limited number of natural videos usually with compression artifacts which are not representative of foveation artifacts and do not have sufficient spatio-temporal variation. Therefore, we collect a new dataset<sup>2</sup> in which we isolate the most relevant factors: velocity, contrast, luminance, and the trade-off between blur and temporal noise (aliasing) due to low sampling rates. The new dataset provides foveation artifacts with sufficiently varied content, and is hence a suitable benchmark to test which quality metrics can correctly account for all these factors.

##### 4.1 Content

We designed a synthetic stimulus containing a uniformly distributed collection of dots moving with controllable velocity, luminance and contrast levels. The stimulus was monochromatic, as the visual system is more sensitive to changes in luminance than chromaticity [Kelly 1983]. Such a synthetic stimulus has several benefits over using photographic content or complex computer graphics scenes. (1) parameters can be precisely controlled, which ensures that the measured content covers a large range of key variables; (2) uniformity ensures that a single salient feature does not distract the observer during trials. We also decided not to use fundamental psychophysical stimuli, such as Gabor patches, as these are not representative of the target content. In our experiments we used two luminance levels ( $Y = \{32.5, 65\}$  cd/m<sup>2</sup>), three contrast levels ( $c = \{0.25, 0.5, 1.0\}$ , Weber contrast), and three different movement speeds ( $v = \{0, 2, 5\}$  visual degrees per second). These values are typical in normal video content, while also fitting within the dynamic range of a modern VR display.

*Foveation.* Foveated rendering can be considered a low-sampling-rate rendering with a non-uniform sampling pattern. We use the sampling rate  $s$  to describe how many samples are considered relative to the full-resolution render target. As discussed in Section 2, there are two key stages of a foveated rendering algorithm: (1)

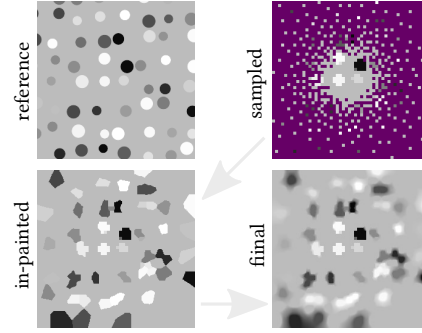


Fig. 11. Illustration of foveation at  $s = 1\%$ . Dark purple color indicates missing samples from the full image with pixels magnified to aid visualization. This is first in-painted with natural neighbors, then box-filtered to produce the rendered image.

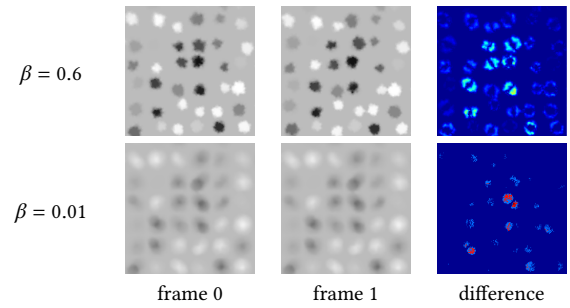


Fig. 12. Flicker and ghosting artifacts on two consecutive frames. When the top sequence is viewed, irregularities in the circular shapes result in perceivable flicker. The bottom sequence is temporally stable but blurry.

sampling in a non-uniform (foveated) manner, and (2) reconstruction of a uniform raster image from the sparse samples. For (1) we use a blue-noise-based sparse sampling pattern, with sample density decreasing with eccentricity in the same fashion as in Deep-Fovea [Kaplanyan et al. 2019]. For reconstruction we selected an in-painting method with natural neighbors (computed in real-time with GPU jump-flooding [Rong and Tan 2006]). To remove sharp edges, the image was then box filtered with a variable filter size equal to the Euclidean distance to the nearest sample at each pixel. Figure 11 illustrates each of these steps.

*Temporal consistency.* As the blue noise pattern has a new seed for each frame, the content is temporally unstable. We employ temporal anti-aliasing with amortized supersampling [Yang et al. 2009] with no reprojection. Each frame  $I$  at time  $t$  is computed by combining the rendered frame  $I_R$  with the history frame  $I_{t-1}$ .

$$I_t = \beta I_R + (1 - \beta)I_{t-1}, \quad (21)$$

where  $\beta$  is the free parameter controlling the trade-off between temporal artifacts (flicker, ghosting) and spatial artifacts (blur). See examples in Figure 12. Once again, we selected this algorithm due to its simplicity and high level of control over spatio-temporal artifacts expected to be present in a generic foveated renderer.

In our experiments we used four sampling rates typical of state-of-the-art foveated rendering algorithms ( $s = \{1, 5, 10, 100\}\%$  with 100%

<sup>2</sup>The dataset is available at: <https://doi.org/10.17863/CAM.68683>

as the reference, and three values of  $\beta$  ( $\{0.01, 0.1, 0.6\}$ ), controlling the trade-off between spatial and temporal artifacts.

This implementation of foveated rendering is inspired by state-of-the-art techniques but it is not meant to compete with their performance; rather it attempts to capture two generally prominent subsampling artifacts — blurring and flickering (temporal aliasing) — in a controllable manner. Low  $\beta$  values result in blur, whereas high  $\beta$  values result in temporal aliasing.

## 4.2 Experiment procedure

We performed a pairwise comparison experiment with a sequential presentation (2IFC) of two animations, 2 seconds each with a one-second gap displaying a blank screen. Each comparison started with the participant pulling the trigger on either controller, which allowed for short rests to be taken between trials. Participants then made their selection by pulling either the left or the right controller triggers. There was no option to rewatch a comparison, and participants did not receive any feedback.

Since testing all combinations of pairs in 5-dimensional space would result in a prohibitively large number of comparisons, we used a block design, in which luminance ( $Y$ ), contrast ( $c$ ), and velocity ( $v$ ) were kept constant in each block and only the sampling rate ( $s$ ) or the temporal trade-off  $\beta$  parameter changed. We further reduced the number of comparisons by presenting only neighboring conditions (e.g. directly comparing  $s = 10\%$  with  $s = 5\%$  and with the reference  $s = 100\%$ , but not with  $s = 1\%$ ).

*Task.* During each 2s clip, participants were asked to maintain fixation on a red fixation cross displayed in the middle of the VR screen. An eye tracker recorded gaze location during the experiment to validate this. Gaze location during the inter-trial gaps were not considered. After both stimuli were played, the task was to select the sequence with higher visual quality, where visual quality was defined as “comfortable to look at”, and “consistent quality within the field of view”. This intended to capture the concept of flicker and blur. Participants received a brief training, where they were familiarized with the content, the controllers, and these two types of artifacts were pointed out to them. Participants were instructed to weigh the two artifacts with their own subjective preference.

*Setup.* We used an HTC Vive Pro Eye with a Unity application for rendering with custom compute shaders for sparse sampling and jump-flooding. A suitably powerful PC was driving the rendering at 90 frames per second. We kept the head position stable with a chinrest.

*Participants.* 35 participants took part in the experiment (14M, 19F, 2 other, aged 25-65, normal or corrected-to-normal vision) and received token compensation. The experiment was authorized by an external institutional review board..

## 4.3 Experiment results

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 JOD units (refer to Section 3.9). Confidence intervals were estimated using bootstrapping.

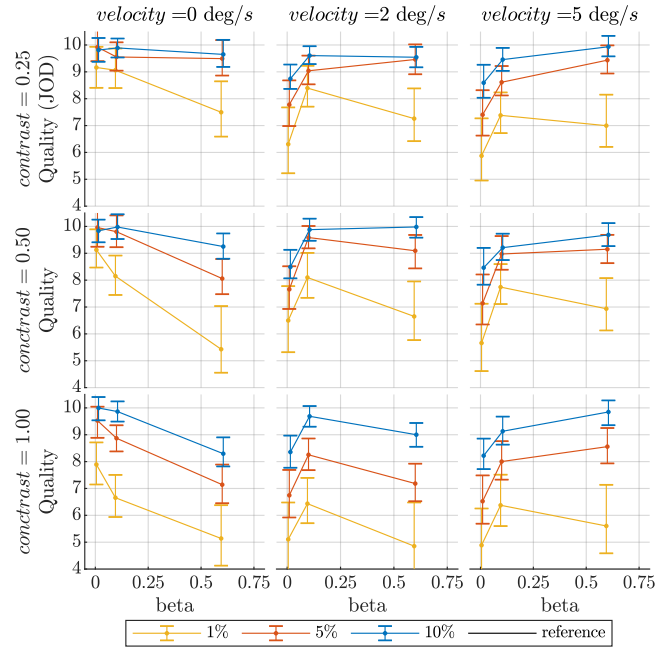


Fig. 13. Results of the psychophysical experiment, showing  $Y = 65 \text{ cd/m}^2$  conditions on linear quality scales. Subplots correspond to all possible contrast ( $c$ ) and motion velocity ( $v$ ) pairing. Colors indicate the sampling percentage ( $s$ ). Error bars denote 95% confidence intervals. The y-axis correspond to subjective quality with 1 JOD difference meaning 50% increase in preference for the method with the higher score.

Figure 13 shows the scaled results for  $65 \text{ cd/m}^2$  for each velocity and contrast pairing as a function of the temporal trade-off factor  $\beta$ . There is a visible difference between the shapes of the curves between different velocities. For stationary images ( $v = 0 \text{ deg/s}$ ), low  $\beta$  values are consistently preferred (strong filtering), as they reduce temporal artifacts and, due to the lack of motion, still produce sharp images. As the velocity increases ( $v = 2 \text{ deg/s}$ ), the trade-off between temporal flicker and spatial blur becomes more apparent with low  $\beta$  values resulting in noticeable motion blur, high  $\beta$  values resulting in objectionable flicker, and preference curves peaking in-between. For high velocities ( $v = 5 \text{ deg/s}$ ), temporal artifacts appear to be less objectionable, and the overall preference is shifted towards higher  $\beta$  values. The contrast levels (rows) follow similar trends, with the differences in quality increased by higher contrast. Luminance (not shown in the figure) did not have a significant impact, most likely due to small relative difference between the two tested luminance levels. Full results including significance tests are available in the supplementary material.

## 5 EVALUATION

A substantial effort was put into testing different variants of the metric and validating the results against several datasets. In this section, we validate the metric on 3 independent foveated video datasets, consisting of a total of 420 video pairs, and on 4 combined image datasets consisting of 4159 SDR and HDR images.

## 5.1 Datasets

In addition to our new foveated rendering dataset, which we will call *FovDots*, we used the following datasets:

*UPIQ*. Unified Photometric Image Quality dataset<sup>3</sup> [Mikhailiuk et al. 2021] consists of over 3779 of SDR and 380 HDR image pairs. The dataset was created by aligning and rescaling quality scores from 2 SDR and 2 HDR datasets: TID2013 [Ponomarenko et al. 2015], LIVE [Sheikh et al. 2006], [Korshunov et al. 2015] and [Narwaria et al. 2013]. We selected this dataset because it contains a large number of test conditions, diversity of artifacts, variation in luminance and dynamic range and most importantly, it has been scaled using the same JOD units as our dataset. This dataset is limited to static images and normal (not foveated) viewing.

*LIVE-FBT-FCVR*. LIVE-Facebook Technologies-Compressed Virtual Reality Databases [Jin et al. 2020, 2019, 2021] consists of the subjective DMOS (differential mean opinion score) of ten 360° videos, distorted using 18 different levels of foveated re-sampling, aggregated over 36 participants. We selected this dataset, as it is one of the few datasets exploring subjective quality of foveated rendering artifacts in wide-field-of-view content.

*DeepFovea*. We used a portion of the dataset used in the validation of DeepFovea [Kaplanyan et al. 2019]. The used portion of the dataset consisted of 78 test videos, compressed with the h.265 video codec in a foveated manner. The video was encoded using different bit-rates in three concentric regions, where the central region was encoded at 50 Mbps and the remaining bit budget was distributed to the two other regions. 57 of the videos had their differential-mean-opinion-scores measured on a large projection screen and 21 videos were tested using HTC Vive Pro HMD. The following frame rates are provided in the dataset: 24, 25, 30 and 60 Hz. We used only the conditions for the foveated h.265 compression method as the reference images and gaze points for other foveated methods were not available to us. For both LIVE-FBT-FCVR and DeepFovea, we modeled the displays used for presentation (HTC Vive Pro or a projection screen) to provide all physically-based metrics with absolute luminance units, correct dimensions and viewing distances.

## 5.2 Calibration protocol

Most quality metrics are usually trained and validated individually on each dataset and their performance is reported as correlation coefficients<sup>4</sup>. We attempted this approach and noted that FovVideoVDP could achieve better correlation coefficients in cross-validation than any other tested metric on each dataset. We could also fit a common set of parameters for all datasets and achieve the highest correlation coefficient among all the metrics we examined. However, when we investigated absolute metric predictions, it was clear that the result was overfitted. It had good performance in establishing quality differences within each dataset but was failing at establishing quality differences across different datasets. This was because correlation coefficients can compensate for the large differences in the content found across the dataset: whether the dataset contains foveated

video or not, whether the effective resolution (in ppp) is high or low, whether the frame-rate was high or low, etc. Therefore, to robustly calibrate our metric and to provide fair validation, it was necessary to calibrate and test against a consolidated dataset, in which quality scores are represented on the same absolute quality scale.

*Dataset merging experiment*. Our goal was to rescale all datasets so that quality value in one dataset reflected the same quality level in another dataset. Both UPIQ and our FovDots datasets represent the quality scores in the same JOD units, so we did not need to perform any alignment of quality scores for them. However, we had to map DMOS values from DeepFovea and MOS values from LIVE-FBT-FCVR to the JOD scale. This is because both MOS and DMOS values use an arbitrary scale, which depends on multiple factors including the training of the participants and the range of distortions present in the dataset. We selected a set of 15 videos from each dataset and asked a panel of 8 experts to rate each video using the JOD scale. To anchor that scale, each expert was provided with a web page with examples of images from the UPIQ dataset at the quality levels of 4, 5, ..., 10 JODs. We asked the observers to match the quality in terms of level of annoyance due to the distortions. We also asked them to watch the videos at full-screen size and from a certain viewing distance, which was calculated depending on the size of their monitor. Because all videos contained foveated content, the experts were asked to assess the video while looking at the fixation point.

After collecting matching quality scores across the datasets, we fitted a linear regression mapping from the native quality scale of each dataset to JODs. Here, we relied on the observation that the relation between DMOS/MOS and JODs is well approximated by a linear function [Perez-Ortiz et al. 2020]. The details of the experiment and mapping procedure are explained in the supplementary.

*Calibration*. In order to calibrate our method, we randomly selected 20% of each dataset for training, leaving the remaining 80% for testing. This proportion allowed the entire 60 GB training set to be memory-mapped for fast random access, and experiments with different proportions did not yield significant improvements. We ran a two-stage optimization process which began with a 2 hour long global optimization pattern search procedure, followed by a gradient-based non-linear constrained optimization using the interior point method, which was allowed to run to convergence. To reduce the number of trained parameters, JOD regression was performed separately after each function evaluation. The loss function used was the RMSE between the metric's prediction and the subjective JOD scores of the merged datasets. We constrained the range of parameter values to lie within a plausible range of psychophysical models to avoid overfitting. As the optimization was performed on a Matlab implementation of the metric, we used numerical differentiation. The optimization of each variant of the metric took between 3 and 8 hours on a cluster node with two Nvidia Tesla P100 GPUs. The values of the optimized parameters are listed in Table 3.

<sup>3</sup>UPIQ dataset: <https://doi.org/10.17863/CAM.62443>.

<sup>4</sup>In addition to the correlation coefficients, RMSE is also commonly reported after individually fitting a non-linear mapping function to each dataset.

Table 3. The parameters of FovVideoVDP (best performing variant).

Model component	Parameters
Contrast sensitivity	$S_{\text{corr}} = 3.1623$ , $\sigma_0 = 1.5$ , $k_{\text{cm}} = 0.4058$
Masking	$k = 0.2854$ , $p = 2.4$ , $q_S = 3.237$ , $q_T = 3.0263$
Pooling	$\beta_x = 0.9575$ , $\beta_b = 1$ , $\beta_c = 0.6848$ , $\beta_f = 1$ , $w_S = 1$ , $w_T = 0.25$
JOD regression	$\alpha_{\text{JOD}} = 0.2495$ , $\beta_{\text{JOD}} = 0.3725$

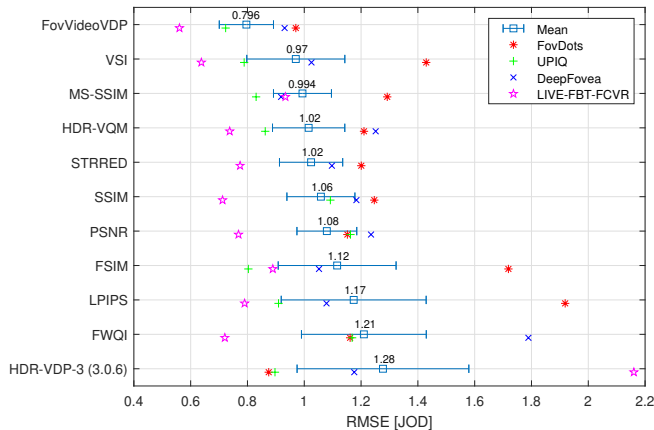


Fig. 14. Comparison of quality metrics in terms of RMSE. The error bars indicate the standard error of the mean.

### 5.3 Comparison with other metrics

We compare the performance of our metric with the subset of relevant metrics from Table 1. The metrics that require physical specifications of the conditions (display size, viewing distance, luminance) are provided with such data. The UPIQ dataset contains 380 HDR images that were calibrated in absolute units of luminance. As traditional metrics cannot operate directly on HDR images, we used the PU21 transform from [Mantiuk and Azimi 2021] to map those images in approximately perceptually uniform units. To find video quality for image metrics, we averaged quality predictions across all frames. For each metric, we fit a non-linearity that maps from metric predictions to the JOD scores of the consolidated datasets. A different non-linearity was selected for each metric. We fit two different non-linearities for HDR-VQM, one for images and another for video, as this metric produces different quality scales for each type of content - note that this could give this metric an undue advantage. The results are reported as RMSE of the prediction in the units of JOD. Because the datasets are imbalanced in terms of number of conditions (4,159 for UPIQ vs. 78 for DeepFovea), we report the average of RMSEs computed per dataset so that each dataset has equal influence on the final average score. We report correlation coefficients (PLCC and SROCC) in the supplementary material, however, note that these are highly affected by the largest dataset (UPIQ), which has no video content, and are less indicative of performance.

The average and per-dataset RMSE of each metric are visualized in Figure 14. FovVideoVDP has a clear lead over the other metrics when the consolidated dataset is considered. The gain in performance is mostly due to its ability to generalize predictions across

the datasets. For example, in the scatter plots in Figure 15, we can see that HDR-VDP-3 does a good job predicting distortions within the UPIQ dataset, but it overpredicts the magnitude of distortions in foveated video datasets, as it does not model foveation or temporal processing. The FWQI metric accounts for foveation and has better accuracy across the datasets, but it has worse precision. Figure 14 shows that most metrics excel in predicting compression distortions found in DeepFovea and LIVE-FBT-FCVR, but struggle with our FovDots dataset containing artifacts due to sampling and filtering.

### 5.4 Ablation and variants

The proposed variant of the metric, described in Section 3, is the winning combination among multiple tested variants. Here, we report the most important findings from testing other possibilities. The results for testing variations of pooling, temporal channels and contrast encoding can be found in the supplementary.

*Masking model.* A masking model is the key component of any metric based on psychophysical models and its selection has a significant impact on the accuracy of the predictions. First, we tested our metric without any masking model, in which perceived difference was encoded as the difference of physical contrast values ( $D_{b,c}(x) = |C_{b,c}^{\text{test}}(x) - C_{b,c}^{\text{ref}}(x)|$ ) or the difference of CSF-normalized contrast values ( $D_{b,c}(x) = |C'_{b,c}{}^{\text{test}}(x) - C'_{b,c}{}^{\text{ref}}(x)|$ ). We also tested the original masking model based on a transducer function, proposed by Foley [1994], and the threshold elevation function used in VDP [Daly 1993]. The equations for those models are included in the supplementary materials. Although we tested multiple variants of these models, in total over 20 variations, these are not included in our analysis as the differences in their performance were too small to select one variant over another. For each variant, we refitted all relevant parameters using our calibration procedure to obtain the best possible performance.

The results shown in Figure 16-top demonstrate that the two variants without the masking model, contrast difference and CSF-normalized contrast difference, perform much worse than those with any masking model. It is interesting to note that the CSF alone does not improve performance. This is most likely because the near-threshold CSF model cannot predict supra-threshold performance due to contrast constancy [Georgeson and Sullivan 1975]. The differences between the masking models are subtle, however, we found a small advantage of the model from Eq. (15).

*Luminance adaptation.* We tested a number of models of both global and local adaptation. The luminance of local adaptation,  $L_a$ , is used in our metric to compute contrast (Eq. (7)) and also to determine the sensitivity (Eq. (11)). We tested the variants in which: (a) we assumed a global level of adaptation and computed  $L_a$  as a geometric mean of the luminance in the image; (b) we used the level of a Gaussian pyramid as shown in Eq. (7), but using the level  $l$ ,  $l + 1$  or  $l + 2$ ; (c) we implemented the model of local adaptation proposed in [Vangorp et al. 2015].

The results in Figure 16-bottom show that the global adaptation model results significantly degraded performance, especially for the UPIQ dataset, which contains HDR images. However, there is

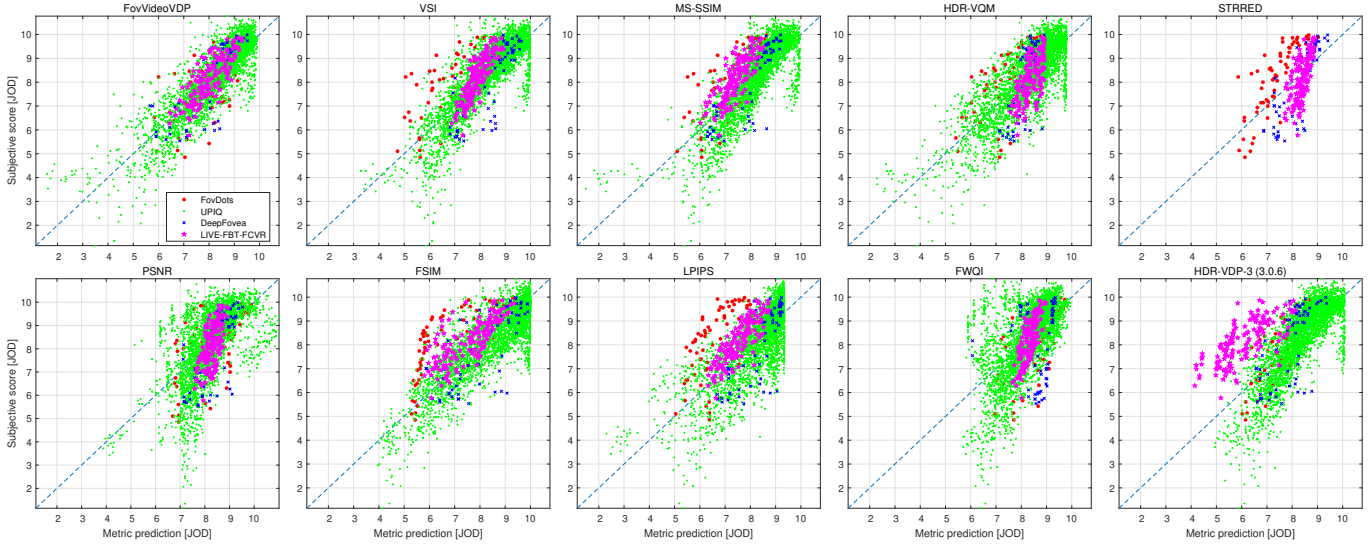


Fig. 15. Subjective vs. predicted quality scores of the compared metrics. Note that STRRED cannot be used to predict the quality of images and is missing predictions for UPIQ.

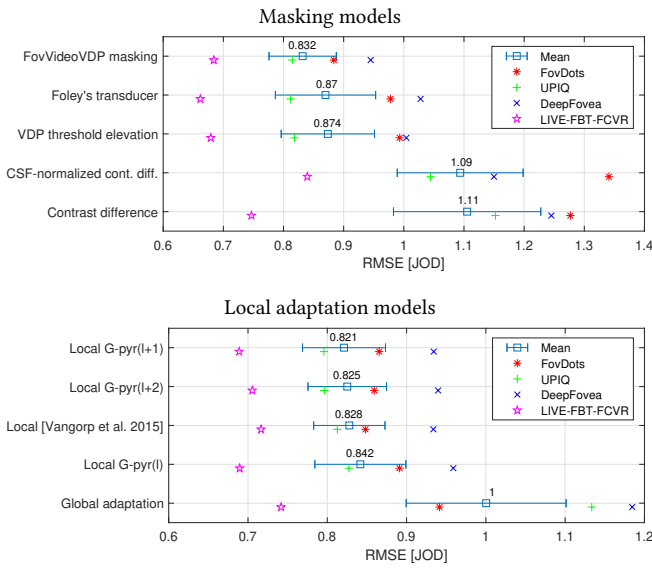


Fig. 16. Ablation studies: *Top*: The variants of the metric that use different masking models. *Bottom*: The variants of the metric that use different local adaptation models.

only a small performance difference between variants of the local adaptation model, and no evidence suggesting that the more complex model of Vangorp et al. [2015] can improve performance. We decided to use the Gaussian pyramid at level  $l + 1$  as a predictor of adapting luminance because it resulted in slightly improved performance. It is readily available as a by-product of constructing the Laplacian pyramid and therefore has a negligible computation cost.

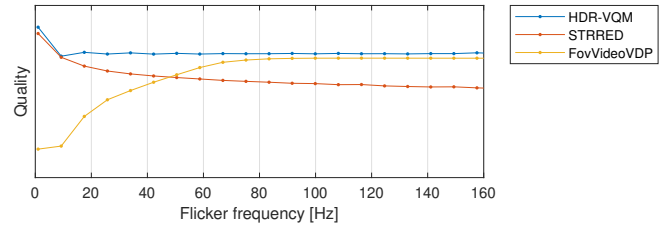


Fig. 17. Metric prediction for flickering square. The quality is arbitrarily scaled for each metric for better comparison.

### 5.5 Synthetic test cases

The limitation of datasets with natural images is that they either confound or do not contain the visual phenomena that a metric models. Therefore, we created a set of 13 synthetic tests, which inspected one aspect of a metric at a time. Here, we report only on flicker and direct the reader to the supplementary for more results.

*Flicker.* We produced 240 Hz video sequences with a square of oscillating luminance with a varying temporal frequencies. The luminance of the background was  $10 \text{ cd/m}^2$  and the contrast of the oscillation was 0.5 (relative to the background). The reference video sequence contained a uniform field (no oscillation). Figure 17 shows the prediction of the flicker visibility for FovVideoVDP and two other metrics, which model the temporal aspects of vision. Our metric correctly predicts that the flicker is fused at higher temporal frequencies and therefore the quality increases with flicker frequency until it saturates (the flicker is fused). The initial drop in quality is also expected as the flicker is the most noticeable at the frequencies of around 5 Hz [Dzn 1952]. Neither of the two other metrics could predict the effect of flicker on quality.

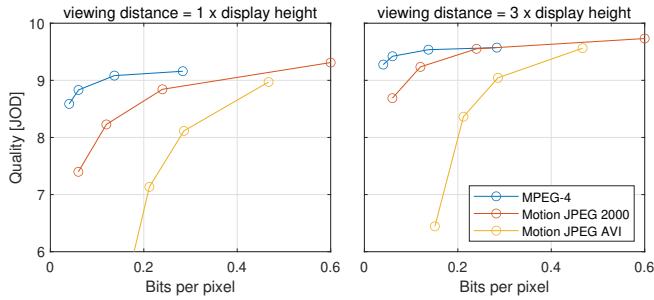


Fig. 18. Comparison of three video compression methods in terms of bit-rate and quality. Since our metric accounts for resolution and screen dimensions, it can be used to predict quality at different viewing distances, specified as multiples of display height (for a 30" display).

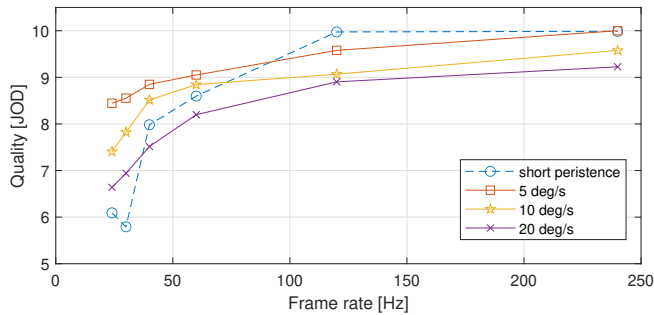


Fig. 19. The predicted quality of motion of a simple target (disk) at different velocities, as seen on displays of different refresh rates. The dashed line shows the quality of the same motion, but seen on a short-persistence display. The quality of a short persistence presentation is the same for all velocities. The reference animation is rendered at 480 Hz.

## 6 APPLICATIONS

In this section we give examples of proof-of-concept applications, which demonstrate the utility of FovVideoVDP .

### 6.1 Video quality assessment

One of the most important applications of video quality metrics is the assessment of lossy video compression methods. FovVideoVDP can be used to assess the quality of SDR or HDR video, both with and without foveated viewing. In Figure 18 we show that it can be used to compare three video codecs at different viewing distances. We can observe that the inter-frame h264/MPEG-4 coding brings the most gains over intra-frame coding (motion JPEG/JPEG 2000) at low bit-rates but also at smaller viewing distances. Such an analysis, which takes viewing distance into account, cannot be performed with the video quality metrics that ignore the physical specification of a display.

### 6.2 Quality of motion

The unique feature of our metric is its ability to assess temporal artifacts, for example those that are due to the limited refresh rate of a display. To illustrate how the quality of motion differs on display of different refresh rates, we render a disk ( $L=80 \text{ cd/m}^2$ ) moving with different velocities on a uniform background ( $L=10 \text{ cd/m}^2$ ). We

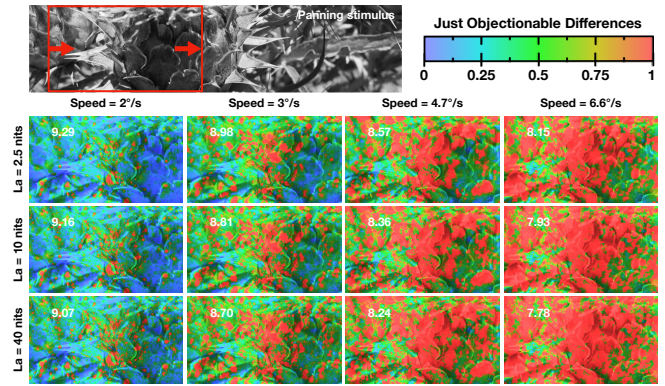


Fig. 20. A panning image stimulus is generated as shown on the top row, at various speeds (2, 3, 4.7, and  $6.6^\circ/s$ ), luminance levels (2.5, 10, and  $40 \text{ cd/m}^2$ ) and framerates (24, 30, and 60Hz). Each video is compared to a sequence with identical luminance and speed at 60Hz. The difference map and final Qjod values (in white) are shown here for the 30Hz case, assuming central fixation. Note the steady increase in the predicted difference with higher speeds and luminance levels.

assume that the eye follows the object perfectly and render the animation as seen by the eye (as projected on the retina). While the gaze moves to follow the disk, an image on the display remains in the same place, creating cyclic motion on the retina, and causing hold-type motion blur. To render such a cyclic motion, we created videos at reference refresh rate of 480 Hz and simulated the motion of the disk on the retina for the given display refresh rate. The metric predictions, shown in Figure 19, indicate that the motion quality drops with lower refresh rates and also the higher velocities of motion. This result is consistent with quality measurements performed on this type of display (see Fig. 7 in [Denes et al. 2020]).

Hold-type motion blur is typically reduced using a short persistence displays, such as those found in VR/AR HMDs. To simulate such a display, we rendered every  $k$ -th frame of 480 Hz video  $k$ -times brighter and set the remaining frames to 0.  $k$  was equal to  $r/480$  where  $r$  was the simulated display refresh rate. Our metric's prediction for such a short-persistence display is shown as a dashed line in Figure 19. It indicates that a short persistence is likely to cause flicker at low refresh rates but it will improve motion quality at higher refresh rates. This prediction holds in practice as short-persistence displays require higher refresh rates to keep the flicker invisible. Such motion quality predictions were previously possible only with specialized models, such as the one proposed by Denes et al. [Denes et al. 2020]. In contrast, our metric is general and makes the predictions based on video content rather than assuming a fixed stimulus (an edge).

Further, we tested our metric as a tool to predict visual distortions due to non-smooth motion artifacts. Chapiro et al. [2019] recently collected data on the perception of judder for scenes with varying luminance levels, speeds and frame rates. While our model is not restricted to predicting motion artifacts, it can be used to detect their presence by comparing videos that vary in frame rate. Fig 20 shows our model operating on the data set used in [Chapiro et al.

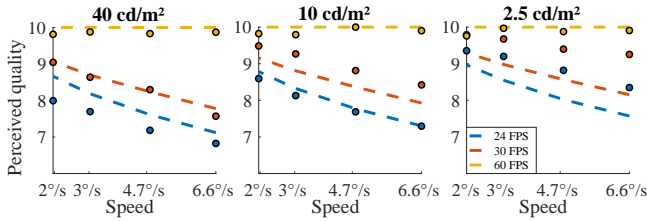


Fig. 21. This plot shows a comparison between our model’s prediction (dashed lines) and the results gathered by Chapiro et al. [2019] (circles) for the panning flower stimulus shown in Figure 20. Our method correctly predicts the trends of increase in perceived judder with higher speeds and decreases with higher frame rates.

2019] *Experiment 3*. Note that our method correctly predicts increasing judder with higher speeds, and similarly a positive trend with increasing luminances. The effect is reduced for higher frame rates, where judder is less visible. Figure 21 contrasts our methods’ predictions to this experimental data, scaled to fit our JOD output.

### 6.3 Foveated sampling

The cost of rendering or transmission of 360 VR animation can be greatly reduced if we can take advantage of gaze-contingent techniques [Kaplanian et al. 2019]. Tursun et al. [2019] noted that the sampling rate can be reduced not only for high eccentricities, but also for low-contrast or low-luminance parts of the scene. We follow a similar approach and predict the lowest sampling rate we can use for 360 video. Such an approach could be used for more efficient transmission of 360 video.

To find a minimum sampling rate, a full resolution frame is sub-sampled (using nearest neighbors) to a fixed set of sampling resolutions: 1x1, 2x2, 4x4 and 8x8. Then, the full resolution frame is reconstructed from such a set of subsampled frames (using nearest-neighbors) and compared with the original using FovVideoVDP. The lowest sampling rate at a particular pixel location is determined to be the one that produces an error of less than 0.25 JOD within each 16x16 tile, according to the difference map. An example of a 360° scene reconstructed using a variable sampling rate is shown together with the sampling map in Figure 22.

### 6.4 Subtle Gaze Direction

Subtle gaze direction [Bailey et al. 2009] applies just-noticeable temporal modulations to peripheral pixels of an image to help direct a user’s gaze without distorting the overall visual experience (Figure 23, top). It is important that the image manipulations are both noticeable and not too distracting for the user. Thus, the ideal stimulus depends on image content as well as viewing conditions. Since FovVideoVDP accounts for spatiotemporal perception and foveation, it can be used to select optimal parameters for the stimulus, avoiding the need for a user study.

We demonstrate this in Figure 23 (bottom), where we simulate subtle gaze direction at various locations of a 1280x1024 image, viewed on a 20-inch monitor from a distance of 75 cm (matching the experimental setup of Bailey et al. [2009]). Using the FovVideoVDP

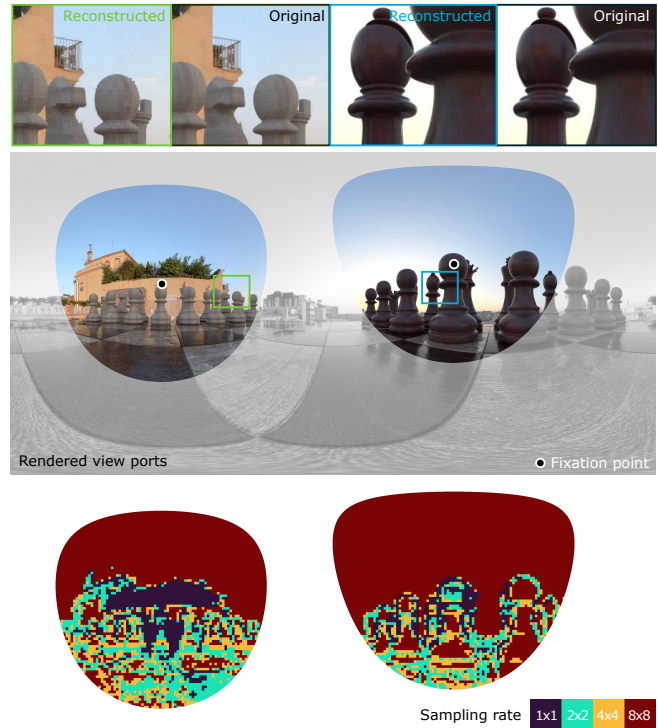


Fig. 22. Two view ports of a 360° video, reconstructed from sparse samples. The video can be sampled at a much lower resolution in the regions of low contrast, low luminance and those far from the gaze location. Note that the regions with uniform textures can be sparsely sampled even if they are close to the gaze location, but the highly textured regions (such as the plants on the building) must be densely sampled. Image courtesy Dabarti CGI Studio and part of dataset accompanying Sitzmann et al. [2017]

output as our objective function, we then find the minimum magnitude of a temporally varying luminance modulation which would result in a difference of at least 0.5 JOD. At a modulation frequency of 10 Hz, we get an average threshold intensity  $i$  of 0.0906, which is close to the value reported (0.095) by Bailey et al. [2009]. Based on the resulting maps, we can also conclude that the optimum stimulus depends largely on the image content, requiring higher strength in high-luminance areas and lower strength in lower-luminance areas. The dependence on eccentricity is not evident, which is expected due to the viewing conditions (narrow field-of-view) as well as the high sensitivity to flicker for both foveal and peripheral viewing. Finally, FovVideoVDP predicts that we require a stronger stimulus for 10 Hz modulation than for a 5 Hz modulation. The latter represents the peak frequency response for the transient channel in our model.

## 7 CONCLUSIONS

In this work we introduce a visual difference metric that is calibrated in physical units, models temporal aspects of vision and accounts for foveated viewing. The main strength of the metric is its ability to generalize across a diverse range of contents and types of spatio-temporal artifacts. This metric was carefully calibrated using 3 independent video quality datasets and a large image quality



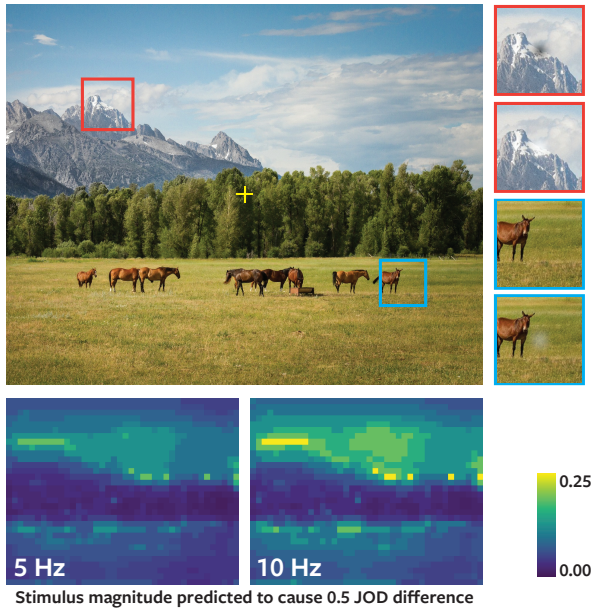


Fig. 23. FovVideoVDP can help optimize parameters of a subtle-gaze direction system. *Top*: Two examples of luminance modulation stimuli at peripheral locations, displayed 75 cm from a 20 inch monitor with user gaze at the center. *Bottom*: Maps showing optimized magnitude (blend factor) for luminance modulation that would result in a difference of 0.5 JOD at various locations across the image. Higher values imply the need for a stronger stimulus. Our metric predicts that high-luminance areas require higher stimulation to result in perceived differences, with the average predicted intensity (0.0906 at 10 Hz) consistent with the value of 0.095 reported by Bailey et al. [2009]. Due to the small field-of-view, the effect of foveation is negligible. Note that a 10 Hz modulation requires a stronger stimulus compared to a 5 Hz one. Image by StockSnap from Pixabay.

dataset. This work demonstrates that a metric founded on the psychophysical models of vision can explain image and video quality well, often out-performing metrics that rely on hand-crafted features, statistical indicators or machine learning. Finally, the metric can be efficiently implemented to run on a GPU, making it one of the fastest metrics of such complexity. The need for such a new metric is driven by the applications in graphics, in particular in AR/VR rendering, that involve foveated viewing or require the assessment of spatio-temporal artifacts, such as blur, flicker and temporal noise. We hope the metric will find its use in optimization and testing of foveated rendering, temporal antialiasing and denoising techniques. In the future, we plan to use the metric as a differentiable loss function, which can be employed to directly optimize both traditional and machine learning techniques. Our metric does not model certain aspect of vision: color, glare, inter-channel masking, and eye motion. We plan to address these extensions in future work.

#### ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement N° 725253-EyeCode).

#### REFERENCES

- Tunç Ozan Aydin, Martin Čadik, Karol Myszkowski, and Hans-Peter Seidel. 2010. Video quality assessment for computer graphics applications. *ACM Transactions on Graphics* 29, 6 (dec 2010), 1. <https://doi.org/10.1145/1882261.1866187>
- Reynold Bailey, Ann McNamara, Nisha Sudarsanam, and Cindy Grimm. 2009. Subtle Gaze Direction. *ACM Transactions on Graphics* 28, 4 (Sept. 2009). <https://doi.org/10.1145/1559755.1559757>
- Peter G. J. Barten. 1999. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE Press. 208 pages.
- Peter G. J. Barten. 2004. Formula for the contrast sensitivity of the human eye. In *Proc. SPIE 5294, Image Quality and System Performance*, Yoichi Miyake and D. Rene Rasmussen (Eds.), 231–238. <https://doi.org/10.1117/12.537476>
- Roy S. Berns. 1996. Methods for characterizing CRT displays. *Displays* 16, 4 (may 1996), 173–182. [https://doi.org/10.1016/0141-9382\(96\)01011-6](https://doi.org/10.1016/0141-9382(96)01011-6)
- Christina A. Burbeck and D. H. Kelly. 1980. Spatiotemporal characteristics of visual mechanisms: excitatory-inhibitory model. *Journal of the Optical Society of America* 70, 9 (sep 1980), 1121. <https://doi.org/10.1364/JOSA.70.001121>
- P. Burt and E. Adelson. 1983. The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications* 31, 4 (apr 1983), 532–540. <https://doi.org/10.1109/TCOM.1983.1095851>
- Alexandre Chapiro, Robin Atkins, and Scott Daly. 2019. A Luminance-Aware Model of Judder Perception. *ACM Transactions on Graphics (TOG)* 38, 5 (2019).
- 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>
- Scott J Daly. 1998. Engineering observations from spatiotemporal and spatiotemporal visual models. In *Human Vision and Electronic Imaging III*, Vol. 3299. International Society for Optics and Photonics, 180–191.
- R.L. De Valois, D.G. Albrecht, and L.G. Thorell. 1982. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research* 22, 5 (1982), 545–559.
- Gyorgy Denes, Akshay Jindal, Aliaksei Mikhailiuk, and Rafal K. Mantiuk. 2020. A perceptual model of motion quality for rendering with adaptive refresh-rate and resolution. *ACM Transactions on Graphics* 39, 4 (jul 2020). <https://doi.org/10.1145/3386569.3392411>
- Robert F. Dougherty, Volker M. Koch, Alyssa A. Brewer, Bernd Fischer, Jan Modersitzki, and Brian A. Wandell. 2003. Visual field representations and locations of visual areas v1/2/3 in human visual cortex. *Journal of Vision* 3, 10 (2003), 586–598. <https://doi.org/10.1167/3.10.1>
- H De Lange Dzn. 1952. Experiments on flicker and some calculations on an electrical analogue of the foveal systems. *Physica* 18, 11 (1952), 935–950.
- J. M. Foley. 1994. Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A* (1994).
- Wilson S. Geisler and Jeffrey S. Perry. 1998. Real-time foveated multiresolution system for low-bandwidth video communication. In *Human Vision and Electronic Imaging III*. SPIE. <https://doi.org/10.1117/12.320120>
- M A Georgeson and G D Sullivan. 1975. Contrast constancy: deblurring in human vision by spatial frequency channels. *J. Physiol.* 252, 3 (nov 1975), 627–656.
- Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D graphics. *ACM Transactions on Graphics* 31, 6 (Nov. 2012), 1. <https://doi.org/10.1145/2366145.2366183>
- S.T. Hammett and A.T. Smith. 1992. Two temporal channels or three? A re-evaluation. *Vision Research* 32, 2 (feb 1992), 285–291. [https://doi.org/10.1016/0042-6989\(92\)90139-A](https://doi.org/10.1016/0042-6989(92)90139-A)
- E Hartmann, B Lachenmayr, and H Brettel. 1979. The peripheral critical flicker frequency. *Vision Research* 19, 9 (1979), 1019–1023.
- Jonathan C. Horton. 1991. The Representation of the Visual Field in Human Striate Cortex. *Archives of Ophthalmology* 109, 6 (June 1991), 816. <https://doi.org/10.1001/archoph.1991.01080060080030>
- Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters* 44, 13 (2008), 800–801.
- Yize Jin, Meixu Chen, Todd Goodall Bell, Zhaolin Wan, and Alan Bovik. 2020. Study of 2D foveated video quality in virtual reality. In *Applications of Digital Image Processing XLIII*, Vol. 11510. International Society for Optics and Photonics, 1151007.
- Yize Jin, Meixu Chen, Todd Goodall, Anjul Patney, and Alan Bovik. 2019. LIVE-Feedback Technologies-Compressed Virtual Reality (LIVE-FBT-FCVR) Databases. <http://live.ece.utexas.edu/research/LIVEFBTFCVR/index.html>.
- Yize Jin, Meixu Chen, Todd Goodall, Anjul Patney, and Alan Bovik. 2021. Subjective and objective quality assessment of 2D and 3D foveated video compression in virtual reality. *IEEE transactions on Image Processing in review* (2021).
- Anton S. Kaplanyan, Anton Sochenov, Thomas Leimkuehler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression using Learned Statistics of Natural Videos. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 38, 4 (2019), 212:1–212:13.
- D. H. Kelly. 1979. Motion and vision II Stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America* 69, 10 (oct 1979), 1340. <https://doi.org/10.1364/JOSA.69.001340>

- D. H. Kelly. 1983. Spatiotemporal variation of chromatic and achromatic contrast thresholds. *JOSA* 73, 6 (1983), 742–750.
- Frederick A.A. Kingdom and Paul Whittle. 1996. Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing. *Vision Research* 36, 6 (1996), 817–829. [https://doi.org/10.1016/0042-6989\(95\)00164-6](https://doi.org/10.1016/0042-6989(95)00164-6)
- Pavel Korshunov, P. Hanhart, T. Richter, A. Artusi, R.K. Mantiuk, and T. Ebrahimi. 2015. Subjective quality assessment database of HDR images compressed with JPEG XT. In *QoMEX*. 1–6. <https://doi.org/10.1109/QoMEX.2015.7148119>
- Justin Laird, Mitchell Rosen, Jeff Pelz, Ethan Montag, and Scott Daly. 2006. Spatio-velocity CSF as a function of retinal velocity using unstabilized stimuli. In *Human Vision and Electronic Imaging*, Vol. 6057. 605705. <https://doi.org/10.1117/12.647870>
- Gordon E. Legge and John M. Foley. 1980. Contrast masking in human vision. *JOSA* 70, 12 (dec 1980), 1458–71.
- Rafal K. Mantiuk and Maryam Azimi. 2021. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *Picture Coding Symposium*.
- Rafal K. Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.* 30, 4, Article 40 (July 2011), 40:1–40:14 pages. <https://doi.org/10.1145/2010324.1964935>
- Rafal K. Mantiuk, Minjung Kim, Maliha Ashraf, Qiang Xu, M. Ronnier Luo, Jasna Martinovic, and Sophie Wuerger. 2020. Practical color contrast sensitivity functions for luminance levels up to 10 000 cd/m<sup>2</sup>. In *Color Imaging Conference*. 1–6. <https://doi.org/10.2352/issn.2169-2629.2020.28.1>
- A. Mikhaliuk, M. Pérez-Ortiz, D. Yue, W. Suen, and R. K. Mantiuk. 2021. Consolidated dataset and metrics for high-dynamic-range image quality. *IEEE Transactions on Multimedia* (2021), (in print).
- Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet, and Romuald Pepion. 2013. Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality. *Optical Engineering* 52, 10 (oct 2013), 102008. <https://doi.org/10.1117/1.OE.52.10.102008>
- Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. 2015. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication* 35 (jul 2015), 46–60. <https://doi.org/10.1016/j.image.2015.04.009>
- Anjul Patney, Marco Salvi, JooHwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 179.
- Eli Peli. 1990. Contrast in complex images. *Journal of the Optical Society of America A* 7, 10 (oct 1990), 2032–2040. <https://doi.org/10.1364/JOSAA.7.002032>
- Eli Peli, Jian Yang, and Robert B. Goldstein. 1991. Image invariance with changes in size: the role of peripheral contrast thresholds. *Journal of the Optical Society of America A* 8, 11 (Nov. 1991), 1762. <https://doi.org/10.1364/josaa.8.001762>
- 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>
- Maria Perez-Ortiz, Aliaksei Mikhaliuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafal K. Mantiuk. 2020. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing* 29 (2020), 1139–1151. <https://doi.org/10.1109/tip.2019.2936103>
- Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Comm.* 30 (2015), 57–77. <https://doi.org/10.1016/j.image.2014.10.009>
- Snježana Rimac-Drlje, Goran Martinović, and Branka Zovko-Cihlar. 2011. Foveation-based content Adaptive Structural Similarity index. In *2011 18th International Conference on Systems, Signals and Image Processing*. IEEE, 1–4.
- Snježana Rimac-Drlje, Mario Vranješ, and Drago Žagar. 2010. Foveated mean squared error—a novel video quality metric. *Multimedia tools and applications* 49, 3 (2010), 425–445.
- J.G. Robson and Norma Graham. 1981. Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research* 21, 3 (jan 1981), 409–418. [https://doi.org/10.1016/0042-6989\(81\)90169-3](https://doi.org/10.1016/0042-6989(81)90169-3)
- Guodong Rong and Tiow-Seng Tan. 2006. Jump flooding in GPU with applications to Voronoi diagram and distance transform. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*. ACM, 109–116.
- J. Rovamo and V. Virsu. 1979. An estimation and application of the human cortical magnification factor. *Experimental Brain Research* 37, 3 (1979), 495–510. <https://doi.org/10.1007/BF00236819>
- Kalpna Seshadrinathan and Alan Conrad Bovik. 2009. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE transactions on image processing* 19, 2 (2009), 335–350.
- H.R. Sheikh, M.F. Sabir, and A.C. Bovik. 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing* 15, 11 (2006), 3440–3451. <https://doi.org/10.1109/TIP.2006.881959>
- E.P. Simoncelli and W.T. Freeman. 2002. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *IEEE ICIP*, Vol. 3. 444–447. <https://doi.org/10.1109/ICIP.1995.537667>
- Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2017. How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics* (2017).
- Philip L. Smith. 1998. Bloch’s law predictions from diffusion process models of detection. *Australian Journal of Psychology* 50, 3 (dec 1998), 139–147. <https://doi.org/10.1080/00049539808258790>
- Rajiv Soundararajan and Alan C Bovik. 2012. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology* 23, 4 (2012), 684–694.
- Srinivas Sridharan, Reynold Bailey, Ann McNamara, and Cindy Grimm. 2012. Subtle gaze manipulation for improved mammography training. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 75–82.
- C. F. Stromeyer and B. Julesz. 1972. Spatial-Frequency Masking in Vision: Critical Bands and Spread of Masking. *Journal of the Optical Society of America* 62, 10 (oct 1972), 1221. <https://doi.org/10.1364/JOSA.62.001221>
- Qi Sun, A. Patney, L.-Y. Wei, O. Shapira, J. Lu, P. Asente, S. Zhu, M. McGuire, D. Luebke, and A. Kaufman. 2018. Towards virtual reality infinite walking: Dynamic saccadic redirection. *ACM Trans. on Graph.* (2018), 16.
- Nicholas T. Swafford, José A. Iglesias-Guitian, Charalampos Koniaris, Bochang Moon, Darren Cosker, and Kenny Mitchell. 2016. User, metric, and computational evaluation of foveated rendering methods. In *Proceedings of the ACM Symposium on Applied Perception - SAP '16*. ACM Press. <https://doi.org/10.1145/2931002.2931011>
- Okan Tarhan Tursun, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. 2019. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics* 38, 4 (July 2019), 1–14. <https://doi.org/10.1145/3306346.3322985>
- Peter Vangorp, Karol Myszkowski, Erich W. Graf, and Rafal K. Mantiuk. 2015. A model of local adaptation. *ACM Transactions on Graphics* 34, 6 (oct 2015), 1–13. <https://doi.org/10.1145/2816795.2818086>
- V. Virsu and J. Rovamo. 1979. Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research* 37, 3 (Nov. 1979). <https://doi.org/10.1007/bf00236818>
- Zhou Wang, Alan C. Bovik, Ligang Lu, and Jack L. Koulouheris. 2001. Foveated wavelet image quality index. In *Applications of Digital Image Processing XXIV*, Andrew G. Tescher (Ed.). SPIE. <https://doi.org/10.1117/12.449797>
- Z Wang, E.P. Simoncelli, and A.C. Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. IEEE, 1398–1402. <https://doi.org/10.1109/ACSSC.2003.1292216>
- 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.
- Andrew B. Watson. 1987. The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing* 39, 3 (sep 1987), 311–327. [https://doi.org/10.1016/S0734-189X\(87\)80184-6](https://doi.org/10.1016/S0734-189X(87)80184-6)
- Andrew B. Watson and Albert J. Ahumada. 2016. The pyramid of visibility. *Human Vision and Electronic Imaging 2016, HVEI 2016* (2016), 37–42. <https://doi.org/10.2352/ISSN.2470-1173.2016.16HVEI-102>
- Stefan Winkler, Murat Kunt, and Christian J van den Branden Lambrecht. 2001. Vision and video: models and applications. In *Vision Models and Applications to Image and Video Processing*. Springer, 201–229.
- Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafal K. Mantiuk. 2018. Dataset and Metrics for Predicting Local Visible Differences. *ACM Transactions on Graphics* 37, 5 (nov 2018), 1–14. <https://doi.org/10.1145/3196493>
- Lei Yang, Diego Nehab, Pedro V. Sander, Pitchaya Sitthi-amorn, Jason Lawrence, and Hugues Hoppe. 2009. Amortized Supersampling. *ACM Trans. Graph.* 28, 5, Article 135 (Dec. 2009), 12 pages. <https://doi.org/10.1145/1618452.1618481>
- Nanyang Ye, Krzysztof Wolski, and Rafal K. Mantiuk. 2019. Predicting Visible Image Differences Under Varying Display Brightness and Viewing Distance. In *CVPR*. 5429–5437. <https://doi.org/10.1109/CVPR.2019.00558>
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. 586–595. <https://doi.org/10.1109/CVPR.2018.00068>

## A PSEUDOCODE

Algorithm 1 provides pseudocode for computing FovVideoVDP given two video sequences, a display model, and model parameters. The pseudocode loosely follows our Pytorch implementation, but we have omitted several optimizations in the interest of clarity. For brevity we have also omitted pseudocode for generating per-pixel difference maps, and direct the reader to Section 3.9.

Algorithm 1. Pseudocode for FovVideoVDP . Some optimizations, e.g. batch computation of Laplacian pyramids, are omitted for clarity.

```

def compute_metric(T_vid, R_vid, display_model, opt):
    # T_vid and R_vid are test and reference video tensors of shape
    # N (number of frames), H (frame height), W (frame width), C (channels=1)
    N = R_vid.shape[0]

    # F contains temporal filter kernels, and Omega contains the peak frequencies of each filter
    F, omega = get_temporal_filters(opt.fps, opt.filter_len) # Eqs. 3 and 4

    for ff in range(N): # For each frame
        R_gpyr = None
        for cc in range(len(F)): # For each temporal channel
            # apply current temporal filter at frame ff
            T_c = apply_temporal_filter(T_vid, frame=ff, F[cc])
            R_c = apply_temporal_filter(R_vid, frame=ff, F[cc])

            # perform Laplacian decomposition
            T_lpyr = decompose_laplacian(T_c)
            R_lpyr = decompose_laplacian(R_c)

            if cc == 0: # perform Gaussian decomposition for sustained channel of the reference
                R_gpyr = decompose_gaussian(R_c)

            for bb in range(R_lpyr[cc].band_count()-1): # For each band, except the base band
                T_f = T_lpyr[cc].get_band(bb)
                R_f = R_lpyr[cc].get_band(bb)

                L_a = R_gpyr.get_band(bb+1) # Local adaptation
                T_con = compute_local_contrast(T_f, L_a) # Eq. 7
                R_con = compute_local_contrast(R_f, L_a) # Eq. 7

                if opt.foveated:
                    ecc = display_model.get_eccentricity(R_con.shape) # compute per-pixel eccentricity
                    res_mag = display_model.get_resolution_magnification(ecc) # compute per-pixel resolution magnification
                else:
                    ecc = zeros(R_con.shape)
                    res_mag = ones(R_con.shape)

                rho = R_lpyr.get_frequencies()[bb] * res_mag # The peak frequency of the band times
                    # the angular resol. magnification (sec 3.1)
                S = compute_contrast_sensitivity(omega[cc], rho, L_a, ecc)
                D = apply_masking_model(T_con, R_con, S)

                Q_per_ch[bb,cc,ff] = p_norm(D.flatten(), opt.beta, axis=0, normalize=True)

    # Pooling, after all frames are done (Eq. 17)
    Q_sc = p_norm(Q_per_ch, opt.beta_sch, axis=0, normalize=False) # pooling across Laplacian bands
    Q_tc = p_norm(Q_sc, opt.beta_tch, axis=1, normalize=False) # pooling across temporal channels[]
    Q = p_norm(Q_tc, opt.beta_t, axis=2, normalize=True) # normalized pooling across time

    # JOD regression (Eq. 19)
    Q_jod = 10.0 + opt.jod_a * pow(Q, opt.beta_jod)

    return Q_jod

```