Adapting Quality Metrics to Tone Mapping

KENNETH CHEN*, New York University, USA
DONGYEON KIM, University of Cambridge, United Kingdom
YUTA ASANO, Reality Labs Research, Meta, USA
ALEXANDRE CHAPIRO, Reality Labs Research, Meta, USA
QI SUN, New York University, USA
RAFAŁ K. MANTIUK, University of Cambridge, United Kingdom

Tone mapping evaluation is difficult because of the substantial differences in absolute luminance between high dynamic range (HDR) reference and tone-mapped standard dynamic range (SDR) test content. To address this challenge, we collected a new tone mapping evaluation dataset, focused on fundamental tone mapping operations, and combined it with several existing tone mapping quality assessment datasets. Rather than introducing new specialized metrics designed for tone-mapped content, we instead developed a set of techniques to adapt existing quality metrics for tone mapping quality assessment. Our approach models the photometric differences between HDR reference and SDR test displays for accurate metric predictions. The technique consists of two steps: first, a display model converts display-encoded content to photometric values; second, these values are re-encoded using a perceptual transfer function to map both HDR and tone-mapped images to the same display-encoded color space. We systematically evaluated both general-purpose image and video quality metrics with our adaptations and those specifically designed for tone mapping. With these adjustments, general-purpose metrics perform much better for tone mapping evaluation, consistently outperforming previously established specialized techniques. Additionally, we adapted the ColorVideoVDP metric to be sensitive to absolute luminance changes, resulting in *ColorVideoVDP-tm*, which shows greatly improved performance and accepts photometric values as input. These results highlight the robustness of our adaptation technique and provide an improved protocol to evaluate future tone mapping quality metrics. Our datasets, code, and supplementary results can be found at kenchen10.github.io/projects/tmometric/index.html.

CCS Concepts: • **Computing methodologies** → **Perception**; **Image processing**.

ACM Reference Format:
Kenneth Chen, Dongyeon Kim, Yuta Asano, Alexandre Chapiro, Qi Sun, and Rafał K. Mantiuk. 2026. Adapting Quality Metrics to Tone Mapping. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3799902.3811107

*All data access, collection, and experiments were performed by New York University and University of Cambridge.

Authors' Contact Information: Kenneth Chen, kennychen@nyu.edu, New York University, Brooklyn, New York, USA; Dongyeon Kim, dongyeon.kim93@gmail.com, University of Cambridge, Cambridge, United Kingdom; Yuta Asano, yasano@meta.com, Reality Labs Research, Meta, Redmond, USA; Alexandre Chapiro, alex@chapiro.net, Reality Labs Research, Meta, Sunnyvale, USA; Qi Sun, qisun@nyu.edu, New York University, Brooklyn, USA; Rafał K. Mantiuk, mantiuk@gmail.com, University of Cambridge, Cambridge, United Kingdom.

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA

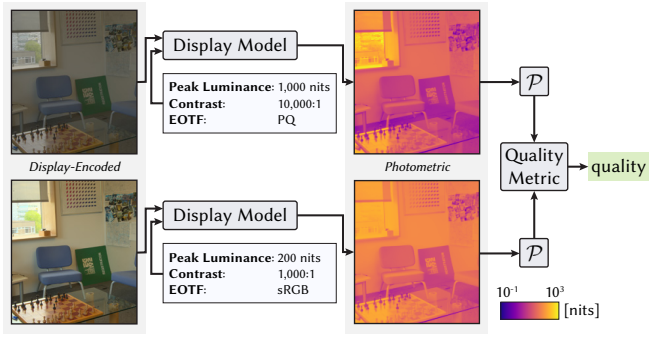
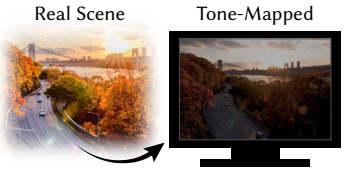


Fig. 1. *Metric adaptation strategy.* The absolute luminance of HDR reference and tone-mapped test can differ significantly. To accurately predict quality, our display model maps display-encoded content to photometric values, given the specifications of the display. The color map represents the photometric values in cd/m^2 (nits). Next, a perceptual transfer function, $\mathcal{P}(\cdot)$, maps these values to a shared display-encoded representation (e.g. 0–1), usable by the quality metric.

1 Introduction



Quality assessment of tone-mapped image and video content is a difficult problem for a number of reasons. Tone mapping introduces complex supra-threshold contrast and color changes that can vary spatially and temporally, or be content-dependent. These factors are often difficult to model perceptually, and are not commonly present in image or video quality datasets for training. Furthermore, the HDR reference is typically scaled *photometrically* (represented in absolute luminance units) while the tone-mapped test is *display-encoded* (e.g., gamma-encoded in a 0–1 range acceptable by general metrics). As a result, traditional metrics are not able to directly compare both types of images accurately.

Subjective studies are the gold standard for quality evaluation, but are costly and infeasible when testing large numbers of combinations, such as those generated by parameter variations in modern tone mapping algorithms. Fully evaluating tone mapping strategies may require an expensive HDR display capable of faithfully reproducing the reference and test images. Given the inevitable changes

introduced by tone mapping, when no HDR reference is available, comparison of tone mapping operators can become affected by individual preference.

Image and video quality metrics can automatically compute quality correlates for tone-mapped content, reducing the need for studies. Most traditional quality metrics, however, are calibrated for SDR content, and their performance is expected to degrade for HDR content. Quality metrics tailored specifically to tone-mapping distortions, by incorporating models of contrast distortion, structural fidelity, or statistical naturalness for instance, exist and are typically calibrated on domain-specific datasets. While such approaches show promising ideas as discussed in Section 2.2, they generally lack quantitative evidence to prove their effectiveness.

In this work, we instead show that correcting for the representational difference between HDR and SDR content can significantly boost the prediction accuracy of *existing* SDR metrics for the task of tone mapping quality assessment. Furthermore, we found that HDR-capable metrics like ColorVideoVDP can be modified to better-account for absolute luminance differences between HDR reference and tone-mapped SDR test content. Our approach requires no re-calibration or additional parameters, and results in a robust evaluation framework that can be applied to existing metric evaluation pipelines for tone-mapped content, leading to consistent improvements over existing specialized metrics.

This is accomplished by our metric adaptation technique, visualized in Figure 1, which first applies a display model to map display-encoded inputs to photometric values. These values are then mapped using a perceptually-uniform transfer function to a shared display-encoded representation. In addition, we developed a modification of ColorVideoVDP [Mantiuk et al. 2024], ColorVideoVDP-tm, that is sensitive to absolute luminance changes by sampling differential contrast sensitivity for HDR reference and SDR test inputs. We evaluated this adaptation strategy and our ColorVideoVDP-tm on several existing tone mapping quality assessment datasets that we gathered, as well as a new tone mapping dataset collected in this work. This evaluation found that encoding both HDR reference and SDR test photometric values using a perceptually-uniform transfer function (PU21) [Mantiuk and Azimi 2021] yields higher performance of adapted general-purpose SDR metrics over specialized tone mapping metrics. In summary, our main contributions are

(1) a subjective tone mapping quality assessment dataset which measures characteristics not explored in prior works[1],
(2) a streamlined methodology to adapt existing quality metrics to the task of tone mapping quality assessment,
(3) a modification to ColorVideoVDP that makes it sensitive to absolute luminance differences between reference and test content (ColorVideoVDP-tm),
(4) and a large-scale evaluation of quality metrics processed with our adaptation strategy across a number of tone mapping quality assessment datasets.

2 Background & Related Work

The HDR reference and SDR test content are represented in different ways. Content captured by a camera or synthesized by a rendering engine is typically scaled linearly with the absolute luminance of the real scene. The cinema community refers to this as *scene-referred* content, because these values are related to the physical light present in the scene. In order to make scene-referred content compatible with the technology and limitations of a specific display, it is converted to a *display-referred* format. In this work, we call scene- and display-referred content *photometric* and *display-encoded* content, respectively. An *opto-electronic transfer function* (OETF, $\mathcal{P}(\cdot)$) maps photometric content to display-encoded values, and its inverse, the *electro-optical transfer function* (EOTF, $\mathcal{P}^{-1}(\cdot)$), maps display-encoded values to linear or photometric values.

Table 1. *Summary of prior experiments.* Here we summarize the prior experiments measuring subjective quality of tone-mapped content. In some studies, users were (✔) or were not (✘) shown the HDR reference, or shown the real scene ((✔)), denoted by the *R* column. A few studies showed users tone-mapped videos (Video column).

Dataset	Video	*R*	# scenes	# conditions
Yeganeh and Wang [2013]	✘	✘	15	120
Drago et al. [2003]	✘	✘	4	24
Kuang et al. [2004]	✘	✘	10	80
Kundu et al. [2017a]	✘	✘	605	1,811
Ak et al. [2023]	✘	✘	250	1,000
Čadík et al. [2008]	✘	(✔)	3	42
Yoshida et al. [2005]	✘	(✔)	2	14
Cerda-Company et al. [2018]	✘	(✔)	3	45
Ledda et al. [2005]	✘	✔	23	138
Melo et al. [2015]	✔	✔	7	42
LUNAM TM Image Quality Dataset [2017]	✘	✔	20	180
Linköping TM HDR Video Dataset [2016]	✔	✘	5	35
LIVE TM HDR Database [2024]	✔	✘	40	1,600
"What is HDR?" [2025]	✔	✔	12	612
Our Dataset	✘	✔	12	420

2.1 How is tone mapping quality measured?

Because HDR content is represented in photometric units linear with absolute luminance (e.g. cd/m^2 or nits), it may contain values outside the range that a typical display can reproduce. Since the development of HDR capture [Debevec and Malik 1997] and display [Seetzen et al. 2004] technology, there has been a proliferation of tone mapping techniques to resolve this issue. Tumblin and Rushmeier [1993] describe the goal of tone mapping as an operation that compresses an HDR scene to the range of the display device, while maintaining its perceptual realism.

A number of works have studied subjective quality of tone-mapped content, with a summary of these in Table 1. We discuss popular tone mappers in the supplement. Most of the studies [Drago et al. 2003; Kuang et al. 2004; Kundu et al. 2017a] asked users to judge tone-mapped images without showing the HDR reference. Others asked users to select tone-mapped content with respect to a real physical scene [Čadík et al. 2008; Yoshida et al. 2005] or with respect to a reference shown on an HDR display [Chen et al. 2025, 2026b; Krasula et al. 2015, 2017; Ledda et al. 2005].

To highlight the differences, we describe a number of these datasets (also shown in Table 1 bottom) in detail below. The following datasets are used in our evaluation of adaptation techniques (see Section 5.2), where L_{max} is display peak luminance, and L_{min} its black level:

[1]Our dataset: https://doi.org/10.17863/CAM.129808

Table 2. *Summary of quality metrics.* This table describes the approach employed by the metric: error-based (e), structural (s), statistical (s), no-reference (n), feature-based (f), video (v), psychophysical (p), and tone mapping (t). R and T represent whether the reference or test input is photometric (✔) or display-encoded (✘), respectively. Metrics with "–" in the R column are no-reference metrics. A more extensive version of this table is displayed in the supplementary document.

Metrics	e	s	s	n	f	v	p	t	R	T
PSNR	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘
CIE ΔE 2000 [CIE 2018]	✔	✘	✘	✘	✘	✘	✘	✘	✔	✔
SSIM [Wang et al. 2004]	✘	✔	✘	✘	✘	✘	✘	✘	✘	✘
BRISQUE [Mittal et al. 2012]	✘	✘	✔	✔	✘	✘	✘	✘	–	✘
DISTS [Ding et al. 2022]	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘
MILO [Çoğalan et al. 2025]	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘
LPIPS [Zhang et al. 2018]	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘
TOPIQ [Chen et al. 2024a]	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘
VMAF [Li et al. 2018]	✘	✔	✘	✘	✘	✔	✘	✘	✘	✘
CGVQM [Jindal et al. 2025]	✘	✘	✘	✘	✔	✔	✘	✘	✘	✘
FLIP [Andersson et al. 2020]	✔	✘	✘	✘	✔	✘	✔	✘	✘	✘
HDR-VDP-3 [Mantiuk et al. 2023]	✘	✘	✘	✘	✘	✘	✔	✘	✔	✔
ColorVideoVDP [Mantiuk et al. 2024]	✘	✘	✘	✘	✘	✔	✔	✘	✔	✔
FFTMI [Krasula et al. 2020]	✘	✔	✘	✘	✘	✘	✘	✔	✔	✘
FSITM [Ziaei Nafchi et al. 2015]	✘	✔	✘	✘	✘	✘	✘	✔	✔	✘
TMQI [Yeganeh and Wang 2013]	✘	✔	✘	✘	✘	✘	✘	✔	✔	✘
CIVDM [Aydin et al. 2008]	✘	✘	✘	✘	✘	✘	✔	✔	✔	✔
ColorVideoVDP-tm	✘	✘	✘	✘	✘	✔	✔	✔	✔	✔

LUNAM TM Image Quality Dataset. Krasula et al. [2017] collected a dataset[2], with 10 natural and 10 synthetic HDR images tone-mapped with 9 popular operators, resulting in 180 tone-mapped images. Participants viewed stimuli on a central HDR reference display ($L_{\max} = 4\,000$ cd/m^2 and $L_{\min} = 0.03$ cd/m^2), with two test SDR displays ($L_{\max} = 200$ cd/m^2) on each side.

Linköping TM HDR Video Dataset. A study of video tone mapping operators was conducted by Eilertsen et al. [2016], consisting of 5 HDR videos tone-mapped with 7 operators, resulting in 35 tone-mapped videos in total. Participants evaluated videos on an SDR display ($L_{\max} = 200$ cd/m^2) with no HDR reference shown.

LIVE TM HDR Video Dataset. Venkataramanan and Bovik [2024] collected a crowd-sourced tone mapping quality database[3] consisting of 15,000 tone-mapped videos from 40 unique HDR videos tone-mapped using 13 operators. They studied several different video compression rates and temporal smoothing strategies.

"What is HDR?". Chen et al. [2025] collected a dataset that consists of 612 HDR videos. 12 HDR reference videos were mapped to different peak luminances and contrasts using an S-shaped curve [Chen et al. 2023]. A single tone curve formulation was used, which makes it a good testbed for whether a metric can predict changes in absolute luminance. Stimuli were shown on a haploscopic HDR display with $L_{\max} = 1\,000$ cd/m^2 and $L_{\min} = 0.001$ cd/m^2.

These differences in setting make evaluation even more difficult, and a display model is required to account for different viewing conditions on photometry.

2.2 Can tone mapping quality be predicted?

Quality metrics specialized to predict tone mapping quality take as input the HDR reference and the tone-mapped test,

$$q = Q_{\mathrm{t}}\left(\mathbf{I}_R,\, \mathbf{D}_T\right), \tag{1}$$

where q is some quality correlate, $Q_{\mathrm{t}}(\cdot)$ is a quality metric (in this case specialized for tone mapping, t), $\mathbf{I}_R$ is the photometric HDR image, and $\mathbf{D}_T$ is the tone-mapped, display-encoded image. The subscripts R and T represent reference and test, respectively. Many of these metrics consider spatial/structural features [Ziaei Nafchi et al. 2015] (s) and measures of "naturalness" or aesthetics computed as the alignment with image statistics [Krasula et al. 2020; Yeganeh and Wang 2013] (s). The contrast independent visual difference metric (CIVDM) by Aydin et al. [2008], built upon HDR-VDP [Mantiuk et al. 2005], was designed to be invariant to contrast changes as long as contrast visibility is preserved. A benefit of this metric is that it can produce maps that localize tone mapping distortions.

Another class of metrics that have been used for tone-mapped content are *no-reference* (n) quality metrics [Cui et al. 2022; Kundu et al. 2017b]. These metrics predict quality given only the tone-mapped test content as input [Mittal et al. 2012, 2013],

$$q = Q_{\mathrm{n}}\left(\mathbf{D}_T\right). \tag{2}$$

Finally, metrics taking photometric input like color difference formulas [CIE 2018] or some error-based (e) measures, and metrics based on psychophysical models (p) (e.g., the visual difference predictors — VDPs [Daly 1992; Mantiuk et al. 2011]) could be used for tone mapping evaluation as they accept both HDR and SDR inputs,

$$q = Q_{\mathrm{p}}\left(\mathbf{I}_R,\, \mathcal{M}_{\mathrm{SDR}}(\mathbf{D}_T)\right). \tag{3}$$

Note that the SDR test must first be converted to photometric values using a display model, $\mathcal{M}_{\mathrm{SDR}}(\cdot)$. In Section 3.1, this display model is described in detail. $Q_{\mathrm{p}}(\cdot,\cdot)$ can represent any metric that accepts photometric inputs.

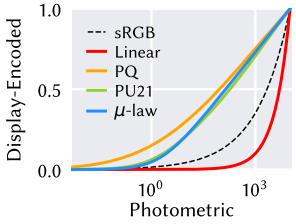
The majority of popular quality metrics are calibrated on general, SDR distortions, and only accept display-encoded content as input,

$$q = Q\left(\mathbf{D}_R,\, \mathbf{D}_T\right). \tag{4}$$

Prior tone mapping quality metrics did not compare against these general metrics. A summary of metrics with their approaches and input types is shown in Table 2.

We show that adaptations can boost prediction quality of metrics not specifically designed for tone mapping quality assessment, even outperforming specialized tone mapping metrics in most cases.

2.3 What is the relation to HDR imaging?



A number of tasks in HDR imaging apply a transfer function (a.k.a. OETF) to map HDR content to a display-encoded, more perceptually-uniform representation, $\mathcal{P}(\mathbf{I}) = \mathbf{D}$. Such tasks include inverse tone mapping [Eilertsen et al. 2017], HDR reconstruction

[2]TMIQD dataset link: sites.google.com/view/lukaskrasula/source-code
[3]LIVE TM HDR Video Dataset: live.ece.utexas.edu/research/LIVE_TMHDR

[Ke et al. 2025], HDR view synthesis [Mildenhall et al. 2022; Xu et al. 2023], HDR image [Mantiuk and Azimi 2021] and video [Venkataramanan et al. 2025] quality assessment. In a similar vein, we show that a naïve strategy to adapt metrics to tone mapping is to apply an OETF to the HDR reference before passing it as input to the metric. Transfer functions considered in this work are shown in the inset (x-axis photometric, y-axis encoded) and described below:

Linear. The simplest strategy is to linearly rescale photometric pixel values to a 0–1 range,

$$\mathcal{P}_{\bigstar}(\mathbf{I}) = \frac{\mathbf{I} - L_{\text{min}}}{L_{\text{max}} - L_{\text{min}}}. \tag{5}$$

However, as shown in the inset, this allocates a very small range of display-encoded values to a wide range of low photometric values.

μ-law. Kalantari and Ramamoorthi [2017] proposed a logarithmic transfer function to encode HDR images for SDR–HDR reconstruction, borrowed from the audio encoding community,

$$\mathcal{P}_{\bigstar}(\mathbf{I}) = \frac{\log(1+\mu\mathbf{I}')}{\log(1+\mu)}, \qquad \mathbf{I}' = \frac{\mathbf{I} - L_{\text{min}}}{L_{\text{max}} - L_{\text{min}}}, \tag{6}$$

where humans were found to have approximately logarithmic perception of loudness, which aligns with Weber-Fechner-like behavior for visual stimuli. Here, we use the value of $\mu = 5,000$.

Perceptual Quantizer. Miller et al. [2013] defined the perceptual quantizer (PQ/SMPTE ST 2084) EOTF which was standardized as an encoding for HDR videos. Its inverse is defined as

$$\mathcal{P}_{\bigstar}(\mathbf{I}) = \left(\frac{c_2\mathbf{I}' + c_1}{1 + c_3\mathbf{I}'}\right)^m, \qquad \mathbf{I}' = \left(\frac{\mathbf{I}}{10,000}\right)^n, \tag{7}$$

where $n = 0.159$, $m = 78.844$, $c_1 = 0.836$, $c_2 = 18.852$, $c_3 = 18.688$. The PQ curve determines perceptually-equivalent steps that map to CSF response; Equation (7) is a fit to that data.

Perceptually-Uniform Transform. Mantiuk and Azimi [2021] defined the PU21 curve, which is similar to PQ but driven by more modern data. The original formulation is fit to a quadratic function,

$$\mathcal{P}_{\bigstar}(\mathbf{I}) = a\left(\log_2(\mathbf{I}) - \log_2(0.005)\right)^2 + b\left(\log_2(\mathbf{I}) - \log_2(0.005)\right), \tag{8}$$

where $a = 0.001908$ and $b = 0.0078$ [Ke et al. 2025]. It was originally developed to map HDR data to a more perceptually-aligned space.

We show below that these transfer functions can also be used for the task of adapting general-purpose image and video quality metrics to tone mapping quality assessment.

3 Method

In this section, we describe the components required to adapt existing metrics to the tone mapping problem. First, we describe how to map display-encoded inputs to photometric values emitted by a display (Section 3.1). Next, we show how to use this display model to adapt existing metrics to predict tone mapping quality (Section 3.2). Finally, we make modifications to ColorVideoVDP that improve its performance on tone mapping evaluation (Section 3.3).

3.1 Display Model

Here, we describe in detail the forward display model, which maps display-encoded to photometric values. The goal of the display model is to convert display-encoded content (e.g., scaled to 0–1) to absolute, photometric values. We use a gain-offset-gamma (GOG) model [Berns 1996] to achieve this,

$$\begin{aligned}\mathbf{I} &= \mathcal{M}_{\text{SDR}}(\mathbf{D}) \\ &= (L_{\text{max}} - L_{\text{min}}) \cdot \mathcal{P}^{-1}(\mathbf{D}) + L_{\text{min}} + L_{\text{refl}},\end{aligned} \tag{9}$$

where L_{refl} is the amount of reflected light. For an SDR display, the EOTF, $\mathcal{P}^{-1}(\cdot)$, may be a gamma function, e.g. sRGB. In the case of HDR content, for example content encoded with the PQ EOTF, display-encoded values are directly converted to photometric values,

$$\begin{aligned}\mathbf{I} &= \mathcal{M}_{\text{HDR}}(\mathbf{D}) \\ &= \mathcal{P}^{-1}(\mathbf{D}).\end{aligned} \tag{10}$$

3.2 Adaptation of Quality Metrics

As we noted in the introduction, image and video quality metrics typically only accept display-encoded inputs. As such, the HDR reference has to be mapped to a range that is compatible with the quality metric, between 0–1 or 0–255 (for 8-bit images) for example. The encodings, $\mathcal{P}(\cdot)$, as described in Section 2.3, achieve this mapping. To compute metric scores, we first map display-encoded inputs to photometric values using the display model,

$$\begin{aligned}\mathbf{I}_T &= \mathcal{M}_{\text{SDR}}(\mathbf{D}_T) \\ \mathbf{I}_R &= \mathcal{M}_{\text{HDR}}(\mathbf{D}_R),\end{aligned} \tag{11}$$

Then, a perceptually uniform encoding is applied to map the photometric test and reference content to a shared display-encoded representation, and then passed as input to a quality metric $Q(\cdot,\cdot)$,

$$q = Q\left(\mathcal{P}(\mathbf{I}_R),\, \mathcal{P}(\mathbf{I}_T)\right). \tag{12}$$

Here, $\mathcal{P}(\cdot)$ represents a perceptually uniform transfer function, which is expected to be better aligned with the response of the human visual system. In this work, we found that applying the PU21 [Mantiuk and Azimi 2021] encoding, $\mathcal{P}_{\bigstar}(\cdot)$, to both the reference and test photometric values yielded the best performance for metric adaptation. A naïve strategy would be to only encode the HDR reference,

$$q = Q\left(\mathcal{P}(\mathbf{I}_R),\, \mathbf{D}_T\right). \tag{13}$$

We tested this strategy with the four transfer functions described in Section 2.3. An extensive discussion of these experiments is described in Section 5.3.

3.3 Adaptation of Visual Difference Predictors

In addition to adaptations to general-purpose SDR metrics, we also found the core reason the VDPs were failing on this task. The VDPs are full-reference video metrics that explicitly model human spatial, temporal and chromatic vision. They operate on photometric values, and, therefore, can directly compare a tone-mapped SDR test content transformed via the display model, Equation (9), with its HDR reference counterpart (see Equation (3)).

We introduce a modification that significantly improves their performance when evaluating tone mapping, demonstrated by creating

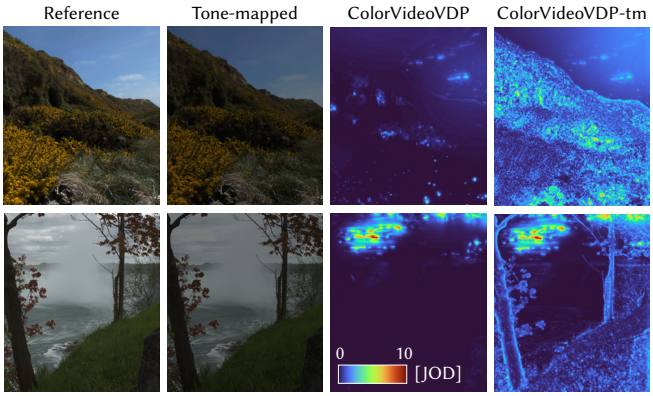


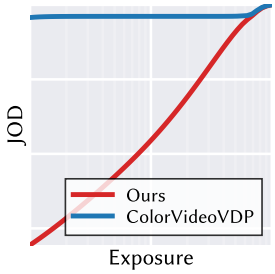
Fig. 2. *Error map visualization.* Our ColorVideoVDP-tm can generate error maps that better account for distortions in absolute luminance compared to the base ColorVideoVDP metric. "Reference" and "Tone-mapped" images here were tone-mapped for presentation.

a variant of ColorVideoVDP with this modification, *ColorVideoVDP-tm*. The metric was originally designed to compare test and reference content at similar luminance, and therefore (for performance reasons) it relied on contrast sensitivity computed for the reference image only.

As shown in the inset (blue line), this makes the metric relatively insensitive to exposure changes, which is a key distortion introduced when tone mapping. We noted that when sensitivity is sampled separately for test and reference content,

$$\begin{aligned} \mathbf{C}'_{R_{(x,y)}} &= \mathbf{C}_{R_{(x,y)}} \mathcal{S}\left(\mathbf{L}_{R_{(x,y)}}\right) \\ \mathbf{C}'_{T_{(x,y)}} &= \mathbf{C}_{T_{(x,y)}} \mathcal{S}\left(\mathbf{L}_{T_{(x,y)}}\right), \end{aligned} \tag{14}$$

the metric becomes much more sensitive to these changes (red line in inset) and its performance on tone mapping datasets is greatly improved. Here, (x, y) is a pixel coordinate, $\mathbf{C}$ is contrast, $\mathbf{L}$ is local background luminance, and $\mathcal{S}(\cdot)$ is contrast sensitivity computed by a contrast sensitivity function (CSF) [Ashraf et al. 2024]. Refer to Mantiuk et al. [2024] (Eq. 7) for the complete formulation.



It should also be stressed that these changes *do not* require re-calibration on any new datasets; ColorVideoVDP-tm uses the same base parameters as ColorVideoVDP, making it backward-compatible. We also found that error maps are improved to visualize changes in absolute luminance due to tone mapping, as shown in Figure 2.

4 Our Dataset

Following the description of our metric adaptation strategy, we present a new tone-mapped image quality assessment dataset collected for this work. The majority of existing tone mapping quality assessment datasets study popular tone mapping operators, and their results determine which operators are preferred. Instead, we focused on a systematic study of the main tone mapping attributes — local and global contrast. Our goals are to test Tumblin's hypothesis on sensitivity to changes in illumination and reflectance (Section 4.2) and to test whether tone mapping metrics can be used to optimize tone mapping parameters (Section 4.3). We developed a generic tone mapper (Section 4.1) to accomplish this, in the spirit of prior tone curve approximation frameworks [Mantiuk and Seidel 2008].

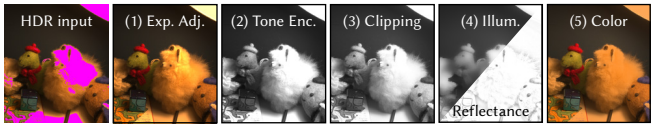


Fig. 3. *Generic tone mapping operator.* Here, we show the operations applied to the HDR "Toys" image in our generic tone mapper. Note that HDR content cannot be accurately shown here, so we visualize out-of-gamut pixels in the input HDR image in magenta.

4.1 Generic Tone Mapping Operator

To generate stimuli for our experiment, we used a simple generic tone mapping operator, with the following components:

(1) exposure adjustment of the HDR image to an SDR range by the p^{th} percentile of the input,

$$\mathbf{I}'_{(x,y)} = \log_2\left(\mathbf{I}_{(x,y)}\right) - \texttt{percentile}\left(\log_2(\mathbf{I}), p\right), \tag{15}$$

(2) tone encoding, $\mathbf{T}$, by applying a cross-channel maximum,

$$\mathbf{T}_{(x,y)} = \max(\mathbf{c}_R, \mathbf{c}_G, \mathbf{c}_B), \quad \mathbf{c} = \mathbf{I}'_{(x,y)}, \tag{16}$$

(3) smooth clipping of values greater than 1 using a spline with control points t_s, t_e, and slope C representing output contrast,

$$\mathbf{T}'_{(x,y)} = \mathcal{G}\left(\mathbf{T}_{(x,y)}, C, t_s, t_e\right), \tag{17}$$

(4) contrast compression (optionally) of illumination ($\mathbf{T}_i$) and reflectance ($\mathbf{T}_r$) components, decomposed using the bilateral filter [Durand and Dorsey 2002], by applying a gamma, γ_i, γ_r, to each component,

$$\mathbf{T}'_{(x,y)} = \gamma_i \cdot \mathbf{T}'_{i_{(x,y)}} + \gamma_r \cdot \mathbf{T}'_{r_{(x,y)}}, \tag{18}$$

(5) and color reconstruction using the method of Mantiuk et al. [2009], with color ratio adjustment s, and sRGB encoded,

$$\mathbf{I}'_{(x,y)} = 2^{\frac{1}{2.2} \cdot \left(s \cdot \left(\mathbf{T}'_{(x,y)} - \mathbf{T}_{(x,y)}\right) + \mathbf{I}_{(x,y)}\right)}. \tag{19}$$

Example outputs from each step are shown in Figure 3, and a more detailed description and diagram are displayed in the supplement. This general framework is used to tone map HDR images in our study, described in the next subsections.

4.2 Illumination vs. Reflectance

Tumblin posits that observers may be more tolerant to lighting (illumination) than detail (reflectance) changes, stating that

> *"Such broad tolerance for lighting changes when making reflectance judgments suggests that the illumination layer of a viewed image or scene is less important and perhaps is sensed less critically than the reflectance layer."*
> — Tumblin et al. [1999]

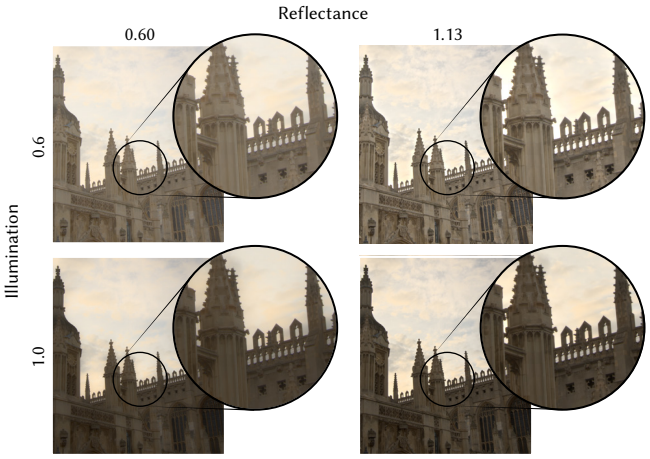


Fig. 4. *Illumination vs. reflectance study stimuli.* A sample of the most extreme parameters in the illumination vs. reflectance study (Section 4.2) are shown here, with increasing reflectance from left to right and increasing illumination from top to bottom. Inset zooms show local detail modification as a result of the bilateral filtering technique [Durand and Dorsey 2002].

This conclusion builds upon experiments by Gilchrist and Jacobsen [1984]. Our goal is to provide data to test this hypothesis in the context of tone mapping of HDR images. We sampled five contrast levels for both the illumination and reflectance layers:

- Illumination (γ_i) = 0.6, 0.7, 0.8, 0.9, 1.0
- Reflectance (γ_r) = 0.6, 0.733, 0.867, 1.0, 1.133

We studied a condition with increased reflectance layer contrast ($\gamma_r = 1.133$). This condition is included because it is a common image enhancement operation and photographers often apply it to maximize detail in photographs. See Figure 4 for example conditions.

4.3 Tone Mapping Optimization

We explored whether tone mapping quality metrics could be used to optimize the parameters of our tone mapping operator. Additionally, we put the tone mapping metrics into direct competition using the maximum absolute differentiation (MAD) competition methodology [Wang and Simoncelli 2008]. In the first case, we found the tone mapping parameters that maximize quality,

$$\begin{aligned} \mathbf{D}_T &= \mathcal{T}(\mathbf{I}_R, \mathbf{k}) \\ \mathbf{k}^* &= \arg\max_{\mathbf{k}} Q_t\left(\mathbf{I}_R, \mathbf{D}_T\right), \end{aligned} \tag{20}$$

where $\mathcal{T}(\mathbf{I}_R, \mathbf{k}) : \mathbf{I}_R \rightarrow \mathbf{D}_T$ is the generic tone mapper described in Section 4.1 and $\mathbf{k}$ its parameters.

Note that we did not use any modifications from Section 3.2, as we used only metrics intended for tone mapping. Khan et al. [2022] conducted a similar procedure to determine failure modes of tone mapping metrics.

In the second case, we maximized the difference in the predictions of a pair of metrics,

$$\mathbf{k}^* = \arg\max_{\mathbf{k}} \left| Q_{t_1}\left(\mathbf{I}_R, \mathbf{D}_T\right) - Q_{t_2}\left(\mathbf{I}_R, \mathbf{D}_T\right) \right|. \tag{21}$$

Since the predictions of each metric do not have the same scale, they were normalized to the range $[0, 1]$ to ensure approximately equal weighting. Optimization was performed by a non-uniform pattern search algorithm [Lewis et al. 2006], and included the image's exposure, contrast, and the tone mapper's smooth clipping range, $\mathbf{k} = \left[p, C, t_e - t_s\right]$. See the supplement for additional details on the optimization procedure. Metrics included TMQI [Yeganeh and Wang 2013], FSITM [Ziaei Nafchi et al. 2015], CIVDM, and ColorVideoVDP (the original, unmodified version). Here, we use $Q_t(\cdot,\cdot)$ to illustrate our procedure, though both CIVDM and ColorVideoVDP accept photometric inputs (i.e. $Q_p(\cdot,\cdot)$, Equation (3)).

4.4 Experimental Methodology

Participants. The study was conducted on 15 participants (10 male, 5 female, aged 19–47). All participants reported normal or corrected-to-normal vision, and were screened for normal color vision using a 10-plate Ishihara test. The participants were compensated for their time and gave informed consent before starting the study. The experiment was approved by the departmental ethics panel.

Hardware Apparatus. An ASUS ProArt PA32UCX 4K HDR mini LED professional monitor[4] was used. The display has a VESA DisplayHDR 1 000 certification[5], is 32 inches across the diagonal, and was calibrated to reproduce a BT.709 color space with PQ response using a JETI specbos 1211UV spectroradiometer[6]. The participants sat at a distance of 0.64 m away from the display, giving an effective number of 60 pixels per degree. The viewing distance was fixed using a chin rest, and room lights were turned off for the duration of the study.

Experimental Procedure. We used a pairwise comparison protocol in which two test images and one reference were shown simultaneously. The reference was an HDR image, which was shown at the center of the display, with two test images shown at either side of the central reference. The test images were tone-mapped versions of the HDR reference, produced using the methods described in Sections 4.2 and 4.3. Participants were asked to select the image "that appears closer to the reference, i.e. the details and colors are more similar to those that can be seen in the reference image". Users interfaced with the study software using a standard keyboard. In order to reduce the number of comparisons, we used active sampling of stimuli via the ASAP algorithm [Mikhailiuk et al. 2021]. ASAP schedules stimuli that maximize expected information gain. The study took on average 21 minutes to complete.

Stimuli. The goal was to select diverse image content, including professional color-graded content, both camera-captured and animated, indoor and outdoor scenes. We also hoped to collect images with details in both bright and dark regions. Images were sourced from the Fairchild [2007] HDR photographic survey, Netflix open source content[7], Stuttgart Media University dataset [Froehlich et al. 2014], and our own set of images captured with a multi-exposure

[4]ASUS HDR display: www.asus.com
[5]DisplayHDR standard: https://displayhdr.org
[6]JETI specbos 1211 spectroradiometer: www.jeti.com
[7]Netflix open source content: https://opencontent.netflix.com

Fig. 5. *User study simuli.* Here, we show the 12 HDR images used in our experiments (tone-mapped for display). See the supplementary document for description of each image.

technique using a Sony α7R III camera. Our captures consist of five exposures fused with the method from Hanji et al. [2020].

We displayed these 12 HDR images to participants, each manually mastered to the peak luminance of the display so that no clipping of highlights occurred. These stimuli are shown in Figure 5 and have resolutions of 1080×944. Tone-mapped images were mapped to simulate an SDR display with 200 cd/m^2 peak luminance and 0.3 cd/m^2 black level. Each study session included both tone mappings described in Sections 4.2 and 4.3, but there were no comparisons between the two. In total, 12 HDR images × (5 γ_i × 5 γ_r + 4 individual metrics + $\binom{4}{2}$ metric combinations) = 420 tone-mapped stimuli were used as conditions in our study. All content was stored as EXR files and rendered in PsychToolbox [Kleiner et al. 2007]. We display all stimuli used in this study in our supplementary webpage.

4.5 Results

We scaled results to just-objectionable-difference (JOD) units using the `pwcmp` algorithm [Perez-Ortiz and Mantiuk 2017]. Higher JOD scores represent greater quality. 1 JOD difference means that 75% of observers selected one condition over another. Outlier detection was computed with the same `pwcmp` package of which none were detected. An N-way analysis of variance (ANOVA) was conducted, which found that the main effect of both illumination (γ_i) and reflectance (γ_r) has a significant effect on JODs ($p \ll 0.01$). The main effect of the metric used in Equations (20) and (21) on JODs was also found to be significant ($p = 0.0001$). All displayed results represent JOD scores averaged across scenes.

Illumination vs. Reflectance. The results in Figure 6 confirm Tumblin's hypothesis that we are more likely to ignore changes in illumination over reflectance. We can see a larger drop in quality and steeper slopes when reflectance contrast is lowered — compare the $\gamma_r = 1$ and $\gamma_i = 1$ curves in both plots. Although this is the case, we are still sensitive to large illumination changes.

Metric Optimization. The results in Figure 7 show that the best tone-mapped image was obtained when the optimization criterion was the FSITM metric and the worst when it was ColorVideoVDP (the original, without the fix from Section 3.3). The worst quality was obtained when CIVDM quality was maximized and FSITM minimized — |CIVDM − FSITM| condition of the MAD analysis.

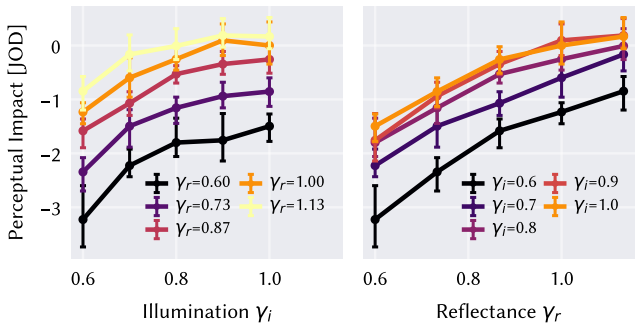


Fig. 6. *Illumination vs. reflectance study data.* The user study data scaled to JODs for the illumination vs. reflectance tone mapper is shown here, with either illumination (left, γ_i) or reflectance (right, γ_r) on the x-axis and JOD on the y-axis. The 0 JOD condition was set to the $\gamma_i = 1, \gamma_r = 1$ condition. Colored lines represent isolated reflectance (left) and illumination (right).

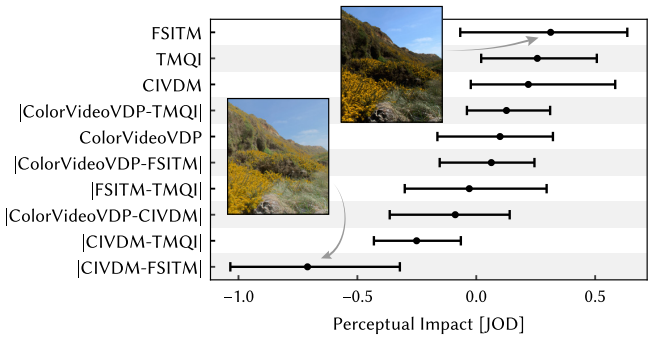


Fig. 7. *Metric optimization study data.* The ranking of different metrics in the metric optimization study is shown, where the y-axis shows quality metrics used for optimization and the x-axis are scaled study results (in JODs, averaged across scenes). We show the lowest and highest-quality tone-mapped images for the "Bloom" scene.

While ColorVideoVDP performed poorly, our metric evaluation in Section 5 found that it performed better on average than the tone mapping metrics at quality prediction. This demonstrates that some metrics may be well-calibrated to perception (good for evaluation) but have poorly-shaped optimization landscapes, and vice versa.

5 Evaluation

Contrary to previous works on tone mapping metrics which compared only with other metrics of their kind, we adapt a large number of state-of-the-art image and video quality metrics for tone mapping evaluation and report their results.

5.1 Protocol

A systematic evaluation of adapted quality metrics, including the improved ColorVideoVDP-tm metric, was conducted across five datasets (Section 5.2) and five adaptation strategies (Section 5.3). We tested a variety of popular quality metrics that accept different input modalities, as discussed in Section 2.2. Our main evaluation metric is the Spearman rank order correlation between subjective

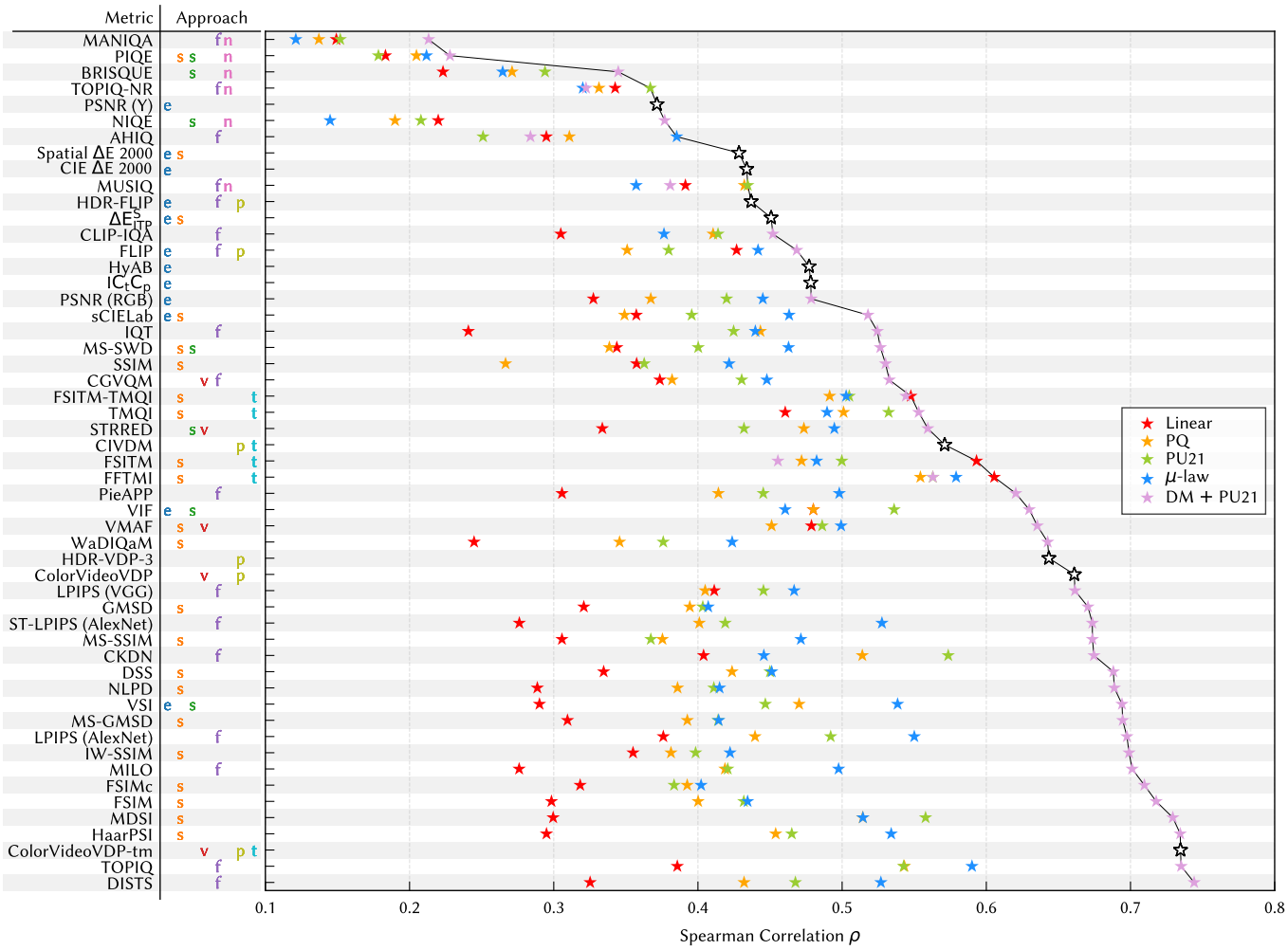


Fig. 8. *Metric evaluation.* Image and video quality metrics are compared with respect to mean Spearman correlation coefficient across the five tone mapping quality assessment datasets. The metrics are sorted (bottom to top) based on correlation across all the display encodings tested, with top-performing encoding techniques connected with a solid line. Colored stars represent different display encoding techniques; white stars represent quality metrics that accept photometric inputs (i.e., do not require display encoding). Note that correlation scores can have a range between 0–1; we plot a constrained range here for better visualization. We also included the metric's approach, with symbols representing error-based (e), structural (s), statistical (s), no-reference (n), feature-based (f), video (v), psychophysical (p), and tone mapping (t) metrics.

and metric-predicted quality scores. In total, our evaluation amounts to (5 adaptation techniques × 5 tone mapping quality assessment datasets × 53 quality metrics) = 1,325 correlation scores.

The results of many of the evaluation datasets are not anchored to the HDR reference quality score. This is either because an HDR display was not available, or it was impossible to construct a JOD scale between an HDR reference and the test conditions (because of unanimous preference for the HDR reference, while the JOD scaling requires disagreement between the observers). As such, quality scores are only relative within each scene. In order to correct for this, we computed the mean correlation across an entire dataset in the space normalized by the Fisher transform,

$$\rho = F^{-1}\left(\frac{1}{N_{\mathcal{D}}}\sum_{d\in\mathcal{D}} F(\rho_d)\right), \quad F(\rho) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{22}$$

where ρ_d is the Spearman correlation for a scene d, and $N_{\mathcal{D}}$ is the number of scenes in the dataset $\mathcal{D}$. The Fisher transform, $F(\cdot)$, makes the correlation values normally distributed and, therefore, allows us to compute the expected correlation value.

5.2 Evaluation Datasets

We evaluated our adaptation strategy on 5 different tone mapping quality assessment datasets. This includes the *LUNAM TM Image Quality Dataset*, *Linköping TM HDR Video Dataset*, *LIVE TM HDR*

Video Dataset, *"What is HDR?"*, and *Our Dataset* (described in Section 4). A subset (1,600/15,000) of the *LIVE TM HDR Video Dataset* was selected in our evaluation; we sampled the lowest compression rate and the "Shotwise" temporal integration. These datasets are summarized in Table 1 (bottom) and described in detail in Section 2.1.

In experimental setups that did not show participants the reference HDR display, we assumed an HDR display model with $L_{\max} = 1\,000\ \mathrm{cd/m^2}$ and $C = 1\,000\,000{:}1$ for our evaluations. In experiments that did not specify the SDR display configurations, we assumed display parameters of $L_{\max} = 200\ \mathrm{cd/m^2}$ and $C = 1\,000{:}1$.

5.3 Ablating Adaptation Strategies

We explore strategies for adapting existing metrics to tone mapping, based on the principles explained in Section 3. Each strategy is denoted by a different star color in Figure 8. The strategies differ in OETF, $\mathcal{P}(\cdot)$, and in the way we represent the tone-mapped image.

We studied four naïve adaptation techniques in which only the HDR reference is encoded (Equation (13)), with either the linear (★), μ-law (★), PQ (★), or PU21 (★) transfer functions. The tone-mapped test is unmodified; for most datasets, this means it is display-encoded in the sRGB color space with value range 0–1. These techniques were compared against our proposed adaptation technique defined in Equation (12), where both HDR reference and SDR test are encoded with PU21 after applying a display model. We call this strategy display model (DM) + PU21 (★).

Note on inputs. In our evaluations, we respected the input types of all metrics. As such, the inputs to tone mapping-specific metrics were unmodified (HDR reference, SDR test). For the *"What is HDR?"* dataset *only*, test inputs were pre-processed using a transfer function to convert them to a display-encoded space; this is because both the test and reference are encoded in an HDR colorspace in this dataset. For all other datasets, tone mapping-specific metrics received unmodified inputs, except when applying DM + PU21 (★) which re-encodes the SDR test with PU21.

5.4 Discussion

The results of our evaluation are plotted in Figure 8, which ranks Spearman correlation scores for the different encoding strategies described in Section 5.3. Each marker in Figure 8 represents a correlation score averaged across all 5 datasets, normalized by the Fisher transform. We also included each metric's approach to better visualize which techniques yielded better performance.

Interestingly, the two top-performing metrics were TOPIQ and DISTS, which are both metrics that use features from deep learning models (f) calibrated on SDR image datasets. Though these two metrics performed well, the performance of deep feature-based metrics was not consistent across the board. The third-best performing metric was our modified ColorVideoVDP, ColorVideoVDP-tm. This metric benefits by being explainable, as it is built off of models of contrast sensitivity. Metrics which accept photometric inputs are agnostic of display encoding; for all encoding strategies except for DM + PU21 (★), these metrics (e.g., ColorVideoVDP-tm, ColorVideoVDP, HDR-VDP-3) performed best. Error-based (e) and no-reference (n) quality metrics consistently underperformed. We also highlight that the metrics built specifically for tone mapping (t) consistently underperformed compared to adapted metrics, with correlations generally in the bottom half of all metrics tested. Per-dataset results (see supplement) across metrics appear well-correlated, with the exception of *Linköping TM HDR Video Dataset*, a small 35-video dataset.

Metric adaptation strategy had a marked impact on correlation scores. When encoding the HDR reference only, the linear encoding (★) consistently performed worse, followed by PQ (★), PU (★), and μ-law (★). The best-performing strategy, by far, was our proposed DM + PU21 (★); ★$<$ ★$<$ ★ $\approx$ ★$\ll$ ★. These results show that bringing both SDR and HDR content to the same color space with the same transfer function is the key to improved performance for tone mapping quality assessment. If only the reference HDR image is encoded, SDR and HDR content are mapped using different OETFs, leading to lower correlations. This highlights the effectiveness of our adaptation strategy for techniques not originally designed for this application.

6 Concluding Remarks

We evaluated our adaptation strategy on an extensive set of quality metrics. The benefit of our technique is that it treats quality metrics as a black box; all operations are computed on *inputs* to a metric, making it easy to adapt any new metric to tone mapping. Our goal was to convey that our metric adaptation strategy is robust, rather than finding the best-performing metric for predicting tone mapping quality. Future work could explore whether our methods may show promise when applied to other applications that result in severe distortions in absolute luminance, such as in augmented reality displays [Chapiro et al. 2024; Kim et al. 2025] or display dimming for power optimization [Chen et al. 2026a, 2024b; Surace et al. 2025].

Conclusion. In this paper, we studied the problem of *adapting* existing general-purpose quality metrics to the task of tone mapping quality assessment. We found that accurate modeling of display photometry is important, and that HDR reference and tone-mapped test content should be encoded using a common perceptually-uniform representation before being passed as input to a quality metric. These adaptations are simple, and produce results better than specialized tone mapping metrics in most cases. In addition, we found a modification to ColorVideoVDP that makes it sensitive to absolute luminance changes, resulting in ColorVideoVDP-tm. Our large-scale evaluation on both existing and our own tone mapping quality assessment datasets shows that our techniques are robust and should be used in future tone mapping evaluation pipelines.

Acknowledgments

Thank you to the user study participants for their time and to the anonymous reviewers for their insightful feedback.

References

Ali Ak, Abhishek Goswami, Wolf Hauser, Patrick Le Callet, and Frederic Dufaux. 2023. RV-TMO: Large-Scale Dataset for Subjective Quality Assessment of Tone Mapped Images. *IEEE Transactions on Multimedia* 25 (2023), 6013–6025. doi:10.1109/TMM.2022.3203211

Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D. Fairchild. 2020. FLIP: A Difference Evaluator for Alternating Images. *Proc. ACM Comput. Graph. Interact. Tech.* 3, 2, Article 15 (Aug. 2020), 23 pages. doi:10.1145/3406183

Maliha Ashraf, Rafał K Mantiuk, Alexandre Chapiro, and Sophie Wuerger. 2024. castleCSF—A contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity. *Journal of vision* 24, 4 (2024), 5–5.

Tunç Ozan Aydin, Rafał Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. 2008. Dynamic range independent image quality assessment. *ACM Trans. Graph.* 27, 3 (Aug. 2008), 1–10. doi:10.1145/1360612.1360668

Roy S. Berns. 1996. Methods for characterizing CRT displays. *Displays* 16, 4 (1996), 173–182. doi:10.1016/0141-9382(96)01011-6 To Achieve WYSIWYG Colour.

Martin Čadík, Michael Wimmer, Laszlo Neumann, and Alessandro Artusi. 2008. Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics* 32, 3 (2008), 330–349.

Xim Cerda-Company, C. Alejandro Parraga, and Xavier Otazu. 2018. Which tone-mapping operator is the best? A comparative study of perceptual quality. *J. Opt. Soc. Am. A* 35, 4 (Apr 2018), 626–638. doi:10.1364/JOSAA.35.000626

Alexandre Chapiro, Dongyeon Kim, Yuta Asano, and Rafał K. Mantiuk. 2024. AR-DAVID: Augmented Reality Display Artifact Video Dataset. *ACM Trans. Graph.* 43, 6, Article 186 (Nov. 2024), 11 pages. doi:10.1145/3687969

Bin Chen, Akshay Jindal, Michal Piovarči, Chao Wang, Hans-Peter Seidel, Piotr Didyk, Karol Myszkowski, Ana Serrano, and Rafał K. Mantiuk. 2023. The effect of display capabilities on the gloss consistency between real and virtual objects. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 90, 11 pages. doi:10.1145/3610548.3618226

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024a. TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment. *IEEE Transactions on Image Processing* 33 (2024), 2404–2418. doi:10.1109/TIP.2024.3378466

Kenneth Chen, Nathan Matsuda, Jon McElvain, Yang Zhao, Thomas Wan, Qi Sun, and Alexandre Chapiro. 2025. What is HDR? Perceptual Impact of Luminance and Contrast in Immersive Displays. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25)*. Association for Computing Machinery, New York, NY, USA, Article 40, 11 pages. doi:10.1145/3721238.3730629

Kenneth Chen, Nathan Matsuda, Thomas Wan, Ajit Ninan, Alexandre Chapiro, and Qi Sun. 2026a. ML-PEA: Machine Learning-Based Perceptual Algorithms for Display Power Optimization. *Computer Graphics Forum* (2026), e70369. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.70369 doi:10.1111/cgf.70369

Kenneth Chen, Thomas Wan, Nathan Matsuda, Ajit Ninan, Alexandre Chapiro, and Qi Sun. 2024b. PEA-PODs: Perceptual Evaluation of Algorithms for Power Optimization in XR Displays. *ACM Trans. Graph.* 43, 4, Article 67 (jul 2024), 17 pages. doi:10.1145/3658126

Kenneth Chen, Yunxiang Zhang, Qi Sun, and Alexandre Chapiro. 2026b. Perceptual Impact of Peak Luminance and Contrast in Direct View HDR Display. *Electronic Imaging* 38, 10 (2026), 222–1–222–1. doi:10.2352/EI.2026.38.10.HVEI-222

CIE. 2018. CIE 015: 2018 Colorimetry. (2018).

Ugur Çoğalan, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Colin Groth. 2025. MILO: A Lightweight Perceptual Quality Metric for Image and Latent-Space Optimization. *ACM Transactions on Graphics (TOG)* 44, 6 (2025). doi:10.1145/3763340

Yueli Cui, Mei Yu, Gangyi Jiang, Zongju Peng, and Fen Chen. 2022. Blind Tone-Mapped HDR Image Quality Measurement by Analysis of Low-Level and High-Level Perceptual Characteristics. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–15. doi:10.1109/TIM.2022.3205928

Scott J. Daly. 1992. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, Bernice E. Rogowitz (Ed.), Vol. 1666. International Society for Optics and Photonics, SPIE, 2 – 15. doi:10.1117/12.135952

Paul E. Debevec and Jitendra Malik. 1997. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 369–378. doi:10.1145/258734.258884

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. 2022. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 5 (2022), 2567–2581. doi:10.1109/TPAMI.2020.3045810

Frédéric Drago, William L. Martens, Karol Myszkowski, and Hans-Peter Seidel. 2003. Perceptual evaluation of tone mapping operators. In *ACM SIGGRAPH 2003 Sketches & Applications* (San Diego, California) *(SIGGRAPH '03)*. Association for Computing Machinery, New York, NY, USA, 1. doi:10.1145/965400.965487

Frédo Durand and Julie Dorsey. 2002. Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (San Antonio, Texas) *(SIGGRAPH '02)*. Association for Computing Machinery, New York, NY, USA, 257–266. doi:10.1145/566570.566574

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM Trans. Graph.* 36, 6, Article 178 (Nov. 2017), 15 pages. doi:10.1145/3130800.3130816

G. Eilertsen, J. Unger, and R.K. Mantiuk. 2016. Chapter 7 - Evaluation of Tone Mapping Operators for HDR Video. In *High Dynamic Range Video*, Frédéric Dufaux, Patrick Le Callet, Rafał K. Mantiuk, and Marta Mrak (Eds.). Academic Press, 185–207. doi:10.1016/B978-0-08-100412-8.00007-3

Mark D. Fairchild. 2007. The HDR Photographic Survey. *Color and Imaging Conference* 15, 1 (2007), 233–233. doi:10.2352/CIC.2007.15.1.art00044

Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. 2014. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In *Digital Photography X*, Nitin Sampat, Radka Tezaur, Sebastiano Battiato, and Boyd A. Fowler (Eds.), Vol. 9023. International Society for Optics and Photonics, SPIE, 90230X. doi:10.1117/12.2040003

Alan Gilchrist and Alan Jacobsen. 1984. Perception of Lightness and Illumination in a World of One Reflectance. *Perception* 13, 1 (1984), 5–19. arXiv:https://doi.org/10.1068/p130005 doi:10.1068/p130005 PMID: 6473052.

Param Hanji, Fangcheng Zhong, and Rafał K. Mantiuk. 2020. Noise-Aware Merging of High Dynamic Range Image Stacks Without Camera Calibration. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, 376–391.

Akshay Jindal, Nabil Sadaka, Manu Mathew Thomas, Anton Sochenov, and Anton Kaplanyan. 2025. CGVQM+D: Computer Graphics Video Quality Metric and Dataset. *Computer Graphics Forum* (2025). doi:10.1111/cgf.70221

Nima Khademi Kalantari and Ravi Ramamoorthi. 2017. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* 36, 4, Article 144 (July 2017), 12 pages. doi:10.1145/3072959.3073609

Andrew Yanzhe Ke, Lei Luo, Xiaoyu Xiang, Yuchen Fan, Rakesh Ranjan, Alexandre Chapiro, and Rafal Mantiuk. 2025. Training Neural Networks on RAW and HDR Images for Restoration Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 810–819.

Ishtiaq Rasool Khan, Theyab A. Alotaibi, Asif Siddiq, and Farid Bourennani. 2022. Evaluating Quantitative Metrics of Tone-Mapped Images. *IEEE Transactions on Image Processing* 31 (2022), 1751–1760. doi:10.1109/TIP.2022.3146640

Dongyeon Kim, Maliha Ashraf, Alexandre Chapiro, and Rafał K. Mantiuk. 2025. Suprathreshold Contrast Perception in Augmented Reality. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25)*. Association for Computing Machinery, New York, NY, USA, Article 68, 11 pages. doi:10.1145/3757377.3763824

Mario Kleiner, David Brainard, and Denis Pelli. 2007. What's new in Psychtoolbox-3? (2007).

Lukáš Krasula, Karel Fliegel, and Patrick Le Callet. 2020. FFTMI: Features Fusion for Natural Tone-Mapped Images Quality Evaluation. *IEEE Transactions on Multimedia* 22, 8 (2020), 2038–2047. doi:10.1109/TMM.2019.2952256

Lukas Krasula, Manish Narwaria, Karel Fliegel, and Patrick Le Callet. 2015. Influence of HDR reference on observers preference in tone-mapped images evaluation. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. 1–6. doi:10.1109/QoMEX.2015.7148103

Lukáš Krasula, Manish Narwaria, Karel Fliegel, and Patrick Le Callet. 2017. Preference of Experience in Image Tone-Mapping: Dataset and Framework for Objective Measures Comparison. *IEEE Journal of Selected Topics in Signal Processing* 11, 1 (2017), 64–74. doi:10.1109/JSTSP.2016.2637168

Jiangtao Kuang, Hiroshi Yamaguchi, Garrett Johnson, and Mark Fairchild. 2004. Testing HDR image rendering algorithms. *Proc. Color Imag. Conf.* (2004), 315–320.

Debarati Kundu, Deepti Ghadiyaram, Alan C. Bovik, and Brian L. Evans. 2017a. Large-Scale Crowdsourced Study for Tone-Mapped HDR Pictures. *IEEE Transactions on Image Processing* 26, 10 (2017), 4725–4740. doi:10.1109/TIP.2017.2713945

Debarati Kundu, Deepti Ghadiyaram, Alan C. Bovik, and Brian L. Evans. 2017b. No-Reference Quality Assessment of Tone-Mapped HDR Pictures. *IEEE Transactions on Image Processing* 26, 6 (2017), 2957–2971. doi:10.1109/TIP.2017.2685941

Patrick Ledda, Alan Chalmers, Tom Troscianko, and Helge Seetzen. 2005. Evaluation of tone mapping operators using a High Dynamic Range display. *ACM Trans. Graph.* 24, 3 (July 2005), 640–648. doi:10.1145/1073204.1073242

Robert Michael Lewis, Virginia Joanne Torczon, and Tamara Gibson Kolda. 2006. *A generating set direct search augmented Lagrangian algorithm for optimization with a combination of general and linear constraints.* Technical Report. Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA ....

Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. 2018. VMAF: The journey continues. *Netflix Technology Blog* 25, 1 (2018).

Rafal Mantiuk, Scott J. Daly, Karol Myszkowski, and Hans-Peter Seidel. 2005. Predicting visible differences in high dynamic range images: model and its calibration. In *Human Vision and Electronic Imaging X*, Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly (Eds.), Vol. 5666. International Society for Optics and Photonics, SPIE, 204 – 214. doi:10.1117/12.586757

Rafał 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), 14 pages. doi:10.1145/2010324.1964935

Rafał Mantiuk, Radosław Mantiuk, Anna Tomaszewska, and Wolfgang Heidrich. 2009. Color correction for tone mapping. *Computer Graphics Forum* 28, 2 (2009), 193–202. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2009.01358.x doi:10.1111/j.1467-8659.2009.01358.x

Rafał Mantiuk and Hans-Peter Seidel. 2008. Modeling a Generic Tone-mapping Operator. *Computer Graphics Forum* 27, 2 (2008), 699–708. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2008.01168.x doi:10.1111/j.1467-8659.2008.01168.x

Rafał K. Mantiuk and Maryam Azimi. 2021. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *2021 Picture Coding Symposium (PCS)*. 1–5. doi:10.1109/PCS50896.2021.9477471

Rafal K Mantiuk, Dounia Hammou, and Param Hanji. 2023. HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625* (2023).

Rafal K. Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. 2024. ColorVideoVDP: A visual difference predictor for image, video and display distortions. *ACM Trans. Graph.* 43, 4, Article 129 (July 2024), 20 pages. doi:10.1145/3658144

M. Melo, M. Bessa, K. Debattista, and A. Chalmers. 2015. Evaluation of Tone-Mapping Operators for HDR Video Under Different Ambient Luminance Levels. *Computer Graphics Forum* 34, 8 (2015), 38–49. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12606 doi:10.1111/cgf.12606

Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafal Mantiuk. 2021. Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization. In *2020 IEEE International Conference on Pattern Recognition (ICPR)*.

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. 2022. NeRF in the Dark: High Dynamic Range View Synthesis From Noisy Raw Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16190–16199.

Scott Miller, Mahdi Nezamabadi, and Scott Daly. 2013. Perceptual signal coding for more efficient usage of bit codes. *SMPTE Motion Imaging Journal* 122, 4 (2013), 52–59.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708. doi:10.1109/TIP.2012.2214050

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* 20, 3 (2013), 209–212. doi:10.1109/LSP.2012.2227726

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

Helge Seetzen, Wolfgang Heidrich, Wolfgang Stuerzlinger, Greg Ward, Lorne Whitehead, Matthew Trentacoste, Abhijeet Ghosh, and Andrejs Vorozcovs. 2004. High dynamic range display systems. *ACM Trans. Graph.* 23, 3 (Aug. 2004), 760–768. doi:10.1145/1015706.1015797

Luca Surace, Jorge Condor, and Piotr Didyk. 2025. Temporal Brightness Management for Immersive Content. (2025). doi:10.2312/sr.20251183

Jack Tumblin, Jessica K. Hodgins, and Brian K. Guenter. 1999. Two methods for display of high contrast images. *ACM Trans. Graph.* 18, 1 (Jan. 1999), 56–94. doi:10.1145/300776.300783

J. Tumblin and H. Rushmeier. 1993. Tone reproduction for realistic images. *IEEE Computer Graphics and Applications* 13, 6 (1993), 42–48. doi:10.1109/38.252554

Abhinau K. Venkataramanan and Alan C. Bovik. 2024. Subjective Quality Assessment of Compressed Tone-Mapped High Dynamic Range Videos. *IEEE Transactions on Image Processing* 33 (2024), 5440–5455. doi:10.1109/TIP.2024.3463418

Abhinau K. Venkataramanan, Cosmin Stejerean, Ioannis Katsavounidis, Hassene Tmar, and Alan C. Bovik. 2025. Cut-FUNQUE: An objective quality model for compressed tone-mapped High Dynamic Range videos. *Signal Processing: Image Communication* 139 (2025), 117405. doi:10.1016/j.image.2025.117405

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

Zhou Wang and Eero P Simoncelli. 2008. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of vision* 8, 12 (2008), 8–8.

Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. 2023. VR-NeRF: High-Fidelity Virtualized Walkable Spaces. In *SIGGRAPH Asia Conference Proceedings*. doi:10.1145/3610548.3618139

Hojatollah Yeganeh and Zhou Wang. 2013. Objective Quality Assessment of Tone-Mapped Images. *IEEE Transactions on Image Processing* 22, 2 (2013), 657–667. doi:10.1109/TIP.2012.2221725

Akiko Yoshida, Volker Blanz, Karol Myszkowski, and Hans-Peter Seidel. 2005. Perceptual evaluation of tone mapping operators with real-world scenes. In *Human Vision and Electronic Imaging X*, Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly (Eds.), Vol. 5666. International Society for Optics and Photonics, SPIE, 192 – 203. doi:10.1117/12.587782

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hossein Ziaei Nafchi, Atena Shahkolaei, Reza Farrahi Moghaddam, and Mohamed Cheriet. 2015. FSITM: A Feature Similarity Index For Tone-Mapped Images. *IEEE Signal Processing Letters* 22, 8 (2015), 1026–1029. doi:10.1109/LSP.2014.2381458