AR-DAVID: Augmented Reality Display Artifact Video Dataset

ALEXANDRE CHAPIRO, Reality Labs, Meta, USA
DONGYEON KIM*, University of Cambridge, UK
YUTA ASANO, Reality Labs, Meta, USA
RAFAŁ K. MANTIUK, University of Cambridge, UK

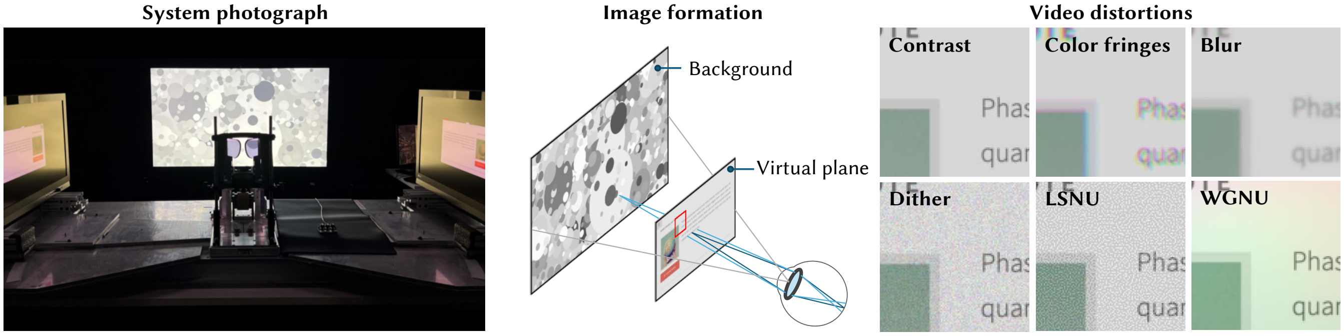

Fig. 1. Our optical-see-through augmented reality testbed (*left*) was built to evaluate the impact of video distortions on perceived quality in the presence of environment light. The background plane is shown with a static pattern, while the virtual plane, positioned closer to user, displays stimulus videos (*center*). Several display distortions were investigated in a pairwise comparison video quality assessment experiment (*right*, enlarged detail of content in the red box).

The perception of visual content in optical-see-through augmented reality (AR) devices is affected by the light coming from the environment. This additional light interacts with the content in a non-trivial manner because of the illusion of transparency, different focal depths, and motion parallax.

To investigate the impact of environment light on display artifact visibility (such as blur or color fringes), we created the first subjective quality dataset targeted toward augmented reality displays. Our study consisted of 6 scenes, each affected by one of 6 distortions at two strength levels, seen against one of 3 background patterns shown at 2 luminance levels: 432 conditions in total. Our dataset shows that environment light has a much smaller masking effect than expected. Further, we show that this effect cannot be explained by compositing of the AR-content with the background using optical blending models. As a consequence, we demonstrate that existing video quality metrics perform worse than expected when predicting the perceived magnitude of degradation in AR displays, motivating further research.

CCS Concepts: • **Hardware → Emerging technologies**.

Additional Key Words and Phrases: augmented reality, visual quality, perception, subjective study

ACM Reference Format:
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 (December 2024), 11 pages. https://doi.org/10.1145/3687969

1 INTRODUCTION

Optical-see-through augmented reality (OST-AR) displays form an additive image, typically at a single focal depth, that blends with light coming from the real world. An image shown on such a display may appear transparent against the background, but this effect is different from naturally occurring transparency, which is typically the result of modulating transmitted light rather than adding new light, as is the case in OST-AR displays. Moreover, there is evidence showing that observers can partially discount either the light coming from the environment or the display [Murdoch 2020; Zhang 2022] when the virtual content is perceived as transparent [Singh and Anderson 2002]. This effect is sometimes associated with veiling luminance, which is discounted to preserve lightness perception [Gilchrist and Jacobsen 1983]. Further, because the environment and display light come from different focal depths and are typically not perfectly aligned, the visual system may gain additional cues, allowing it to partially discard either source of light.

Given the information above, we cannot assume that content seen on AR displays will be perceived in the same way as content shown on traditional displays. Because of this, image/video quality metrics intended for regular display may not perform as expected when used with AR content. To investigate the effect of environment (background) light in AR on quality assessment, we created a new video quality dataset: the Augmented Reality Display Artifact Video Dataset or AR-DAVID[1] Similar to recent work for traditional displays [Mantiuk et al. 2024], AR-DAVID measured the loss of quality due to display distortions, such as blur, contrast loss

*Corresponding author

Authors' addresses: Alexandre Chapiro, alex@chapiro.net, Reality Labs, Meta, Sunnyvale, USA; Dongyeon Kim, dk721@cam.ac.uk, University of Cambridge, Cambridge, UK; Yuta Asano, yasano@meta.com, Reality Labs, Meta, Redmond, USA; Rafał K. Mantiuk, rafal.mantiuk@cl.cam.ac.uk, University of Cambridge, Cambridge, UK.

0730-0301/2024/12-ART186
https://doi.org/10.1145/3687969

[1] Project page: https://www.cl.cam.ac.uk/research/rainbow/projects/ardavid/
Dataset: https://doi.org/10.17863/CAM.111909

due to elevated black level, color fringes, spatiotemporal dithering, light source non-uniformity (LSNU) or waveguide non-uniformity (WGNU) (see the right panel in Fig. 1).

AR-DAVID is the first large-scale visual quality dataset conducted on a custom OST-AR test bed (a haploscope with beam splitters, shown in the left panel on Fig. 1). Distortions were measured for six representative video clips shown over three different background patterns at one of two luminance levels (10 and 100 cd/m^2). In total, the dataset consists of 432 unique distorted videos. 55 users took part, and over 11 000 pairwise comparisons were gathered.

To adapt image and video quality metrics to our AR content, we designed five models simulating the optical blending of the foreground video with the background patterns. We tested 16 state-of-the-art metrics in combination with these blending strategies to examine their accuracy for this new application.

Our main contribution is the AR-DAVID dataset, revealing that the masking effect of the background in OST-AR is much weaker than what would be expected from the optical blending of the light.

2 RELATED WORK

2.1 Perception in OST-AR

The main difference between OST-AR and traditional display is the presence of the visible background, in which real objects are occluded by virtual content shown on the display. In turn, the visibility of the virtual content is affected by the background content. The most straightforward solution to this problem is to make virtual objects much brighter than the real environment so that the contrast of the background is masked following the Weber law. However, this requires virtual objects to be up to 60× brighter than the real scene [Liu et al. 2022]. To achieve such a luminance ratio, the real scene can be dimmed by placing a passive light-attenuating layer (e.g., neutral density filter) on AR glasses. An alternative approach is to block only a part of the real scene, depending on its geometrical relation to the virtual content. However, achieving pixel-wise blocking adaptive to the background can be challenging in terms of the form factor [Gao et al. 2012] of the displays with additional 3D occlusion [Rathinavel et al. 2019] capabilities and latency requirements [Zou et al. 2021]. Instead, software-based adaptive solutions to background intrusion have been introduced for overlaid projection displays [Menk and Koch 2012], focusing on color matching tasks and accelerated by computation-efficient solutions [Hincapié-Ramos et al. 2015]. More recently, Zhang et al. [2021] enhanced the color contrast by optimizing the display pattern with several perceptually-driven constraints. However, these solutions do not solve the occlusion problem, leaving the OST-AR scenes to appear partially transparent.

The human visual response to transparent objects has been shown to be challenging to explain based solely on the physical characteristics of individual objects. This is sometimes referred to as the scission effect [Metelli 1974]. Consequently, this approach has led researchers to modeling AR overlays via non-physical weighted sum of AR foreground and real-scene background. In color matching tasks, Hassani [2019] observed a biased weight on the AR foreground, discounting the effect of the background. However, Murdoch [2020] introduced contradictory results in the brightness matching tasks, suggesting the discounting of foreground weight instead, with dependence on object size. These inconsistent results extended the investigation towards color matching tasks [Zhang 2022] under several luminance conditions, where the weight deviated significantly from unity, particularly with cool background temperatures and low luminance levels. While experimental conditions typically involve flat surfaces at identical depths, natural scenes contain a range of spatial frequencies [Geisler 2008] and depth distributions. The complexity of backgrounds introduces additional dimensions to the AR overlay, making explicit modeling non-trivial.

2.2 AR quality datasets

Most video quality datasets originate from the signal processing community and are meant to capture the effect of streaming or video compression distortions [Min et al. 2024]. Datasets originating from the computer graphics community sometimes target other topics, like geometric distortions [Nehmé et al. 2023; Wolski et al. 2022] or distortions found in foveated rendering [Mantiuk et al. 2021].

In this work, we focus on distortions present in AR displays, such as color fringes or waveguide non-uniformity. By quantifying the impact of this type of distortion on visual quality, we can optimize display design to balance cost, performance, and quality. Closely related work for traditional displays was recently done by Mantiuk et al. [2024]. Their dataset contained 9 display distortion types at 3 strength levels, applied to 14 base video clips. Unlike this work, their measurements were done in controlled conditions on a regular display with no interference from the environment. In our work, content is presented on an OST-AR prototype display, with the virtual content superimposed on backgrounds that simulate aspects of real-world environments. As no video-quality datasets capturing the perception of distortions in OST-AR exist today, our work is meant to fill this gap and enable rigorous display quality research for this display modality going forward.

2.3 Quality metrics for AR displays

When assessing content quality in OST-AR, we need to consider how they differ from regular displays and how this may affect image quality. AR headsets are often equipped with stereoscopic displays. Incorrectly presented stereo content can lead to vergence-accommodation conflict [Koulieris et al. 2017], binocular rivalry [Wang et al. 2024], or induce VR-sickness in the presence of conflicting cues [Eftekharifar et al. 2021]. In this work, we do not consider stereoscopic distortions, but we introduce disparity between the foreground virtual plane and the background environment.

The unique optical properties of OST-AR are typically modeled in linear (photometric/colorimetric) color spaces to preserve physical accuracy. In addition, the blending of light coming from the background can also be modeled in a linear color space. As a consequence, quality metrics for OST-AR may benefit from operating on photometric (or colorimetric) quantities, for example, by taking CIE XYZ trichromatic pixel values as input.

Color difference metrics, such as CIEDE2000 [Sharma et al. 2005] or Delta-ITP [Lu et al. 2016], directly operate on colorimetric quantities but do not model any spatiotemporal aspects of vision. Other

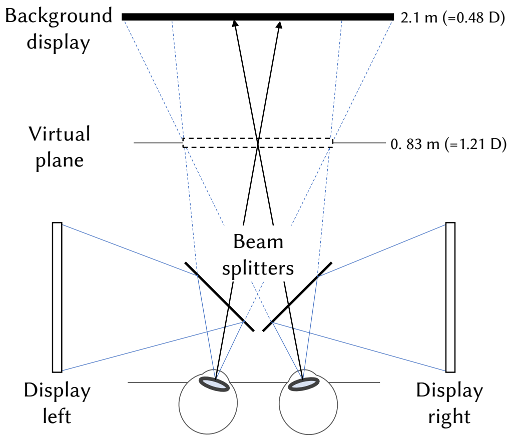

Fig. 2. *The optical arrangement of the haploscope used in our experiment.*

models,such as the visual difference predictors including the original VDP [Daly 1992], HDR-VDP [Mantiuk et al. 2011], FovVideoVDP [Mantiuk et al. 2021] or ColorVideoVDP [Mantiuk et al. 2024], as well as some HDR-capable metrics, including HDR-VQM [Narwaria et al. 2015] model both spatial or temporal detail and photometry. As an alternative, most existing metrics can be adapted to work with photometric quantities by employing a transform function to map input content into a perceptually uniform space [Aydın et al. 2008; Mantiuk and Azimi 2021].

There are no metrics that can explicitly account for the perception of content in OST-AR and, in particular, the effect of the background real-world environment on the visibility of content detail or display-related distortions. This paper is meant to provide the first dataset serving as a starting point for research on this problem.

3 METHOD

3.1 Experiment setup

To create a perceptual dataset of distortions in OST-AR, we built a custom test bed (see photo in Fig. 1 and schematic in Fig. 2).

Virtual plane. The virtual image was produced by two 31.1" Eizo CG3146 professional reference displays, placed to either side of the observer's head. These monitors have a resolution of 4096 × 2160, 60 Hz frame rate, and a contrast value of one-million-to-one. The displays were placed at an effective distance of 83 cm from the observer (1.21 diopters (D)). This produced a field-of-view of $46^\circ \times 26^\circ$. Displayed pixels were replicated in a 2×2 pattern, obtaining an effective resolution of 44.2 pixels-per-degree (ppd). These values were selected to be comparable to commercially available AR devices such as Hololens 2 ($43^\circ \times 29^\circ$ field of view, 33.5 ppd resolution), and Magic Leap 2 ($44^\circ \times 53^\circ$ field of view, 32.7 ppd resolution).

Background plane. The background plane was shown on a 76" DynaScan DK751DH5 screen, placed 210 cm away from the viewer (0.48 D). This distance provided a separation of 0.73 D between planes, which was deemed sufficient to provide clear depth separation. Having the background plane presented via a display was preferred over alternatives (e.g. a poster illuminated by a light source) as a display was easier to calibrate consistently, and allowed for dynamically changing backgrounds between conditions.

Optical path. A pair of beam splitters (Edmunds Optics, neutral response in the visible spectrum) were placed in the user's optical path, oriented at 45 degrees. The virtual plane images are produced via reflection (80%), and the background plane is seen through transmitted light (20%).

Calibration. All three displays in the system were calibrated through the beam splitters using a CS2000-A spectroradiometer. Virtual plane displays were calibrated to a peak luminance of 300 nits, with an sRGB EOTF and P3 color primaries. The background plane display had a peak luminance of 1125 nits, and a gamma of 3.5. All displays were set to a D65 whitepoint. In addition, foreground displays were re-calibrated daily using the built-in colorimeter with a custom externally-loaded calibration matrix to ensure minimal drift throughout the study. The virtual display optical path was also laser-aligned to ensure sub-pixel geometric positioning accuracy.

3.2 Stimuli

The stimuli consisted of a video shown on the foreground virtual plane (simulating an OST-AR display), optically combined with a static pattern shown on the background display (see Fig. 2).

Virtual plane. Content shown on the virtual plane was based on six reference videos showing natural and rendered scenes (see Fig. 3). To provide maximal coverage for plausible AR use, our dataset includes scenes containing both human and animated subjects, high frequencies and flat regions, text, animation, and user interfaces. Three scenes (*Caminandes, Emojis, Foliage*) were selected from the dataset of Mantiuk et al. [2024], while the remaining scenes (*Blog, Messaging, Talking*) were created for this study.

To produce distorted versions of the references, each video was modified using one of six distortions (*blur, color fringes, contrast loss, dither, light source nonuniformity, and waveguide nonuniformity*, shown in Fig. 4). These artifacts were produced in the same way as detailed by Mantiuk et al. [2024]. As the background present in an AR display is expected to mask the visibility of distortions, only the two higher distortion levels (2 and 3) were included in this study.

To avoid excessive study size, two distortions presented by Mantiuk et al. [2024] were excluded as they produced very subtle metric responses which could be expected to become invisible in an AR scenario: *Chroma Subsampling* and *Dynamic Correction Error.*

Background content. The background content was formed by static images (see Fig. 3). These were presented at two mean luminance levels: 10 (dim) and 100 (bright) cd/m^2. These luminance levels are meant to model plausible brightness values present in indoor scenes [Matsuda et al. 2022]. Three different patterns were employed:

- A flat image was used to represent simple backgrounds, like a featureless wall.
- A pink noise pattern ($1/f^x$ power spectrum with $x \approx 1.8$) was found to have a similar frequency profile to that of natural images [Ruderman and Bialek 1993; Tolhurst et al. 1992].

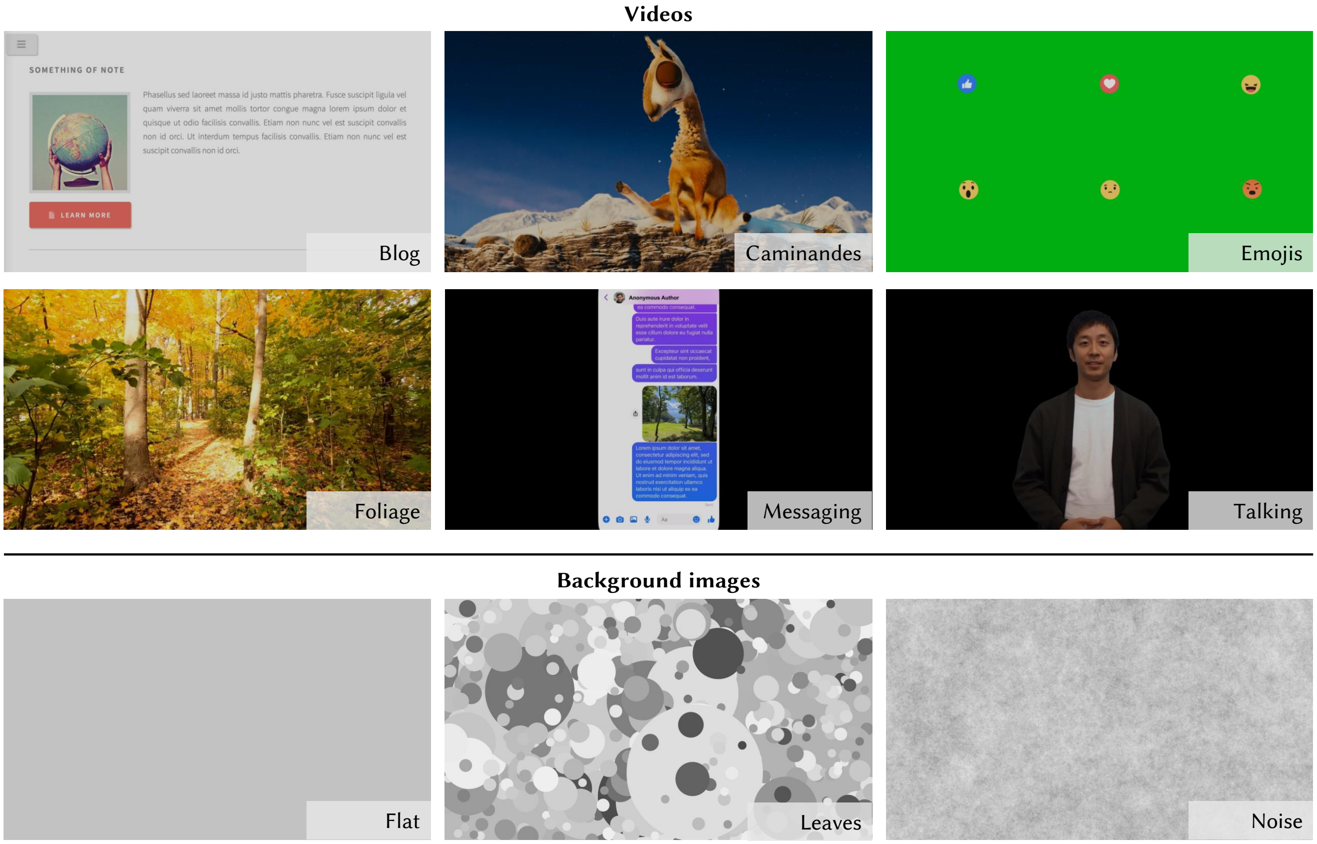

Fig. 3. *Six reference video sequences (top) and three backgrounds (bottom) used to create AR-DAVID.*

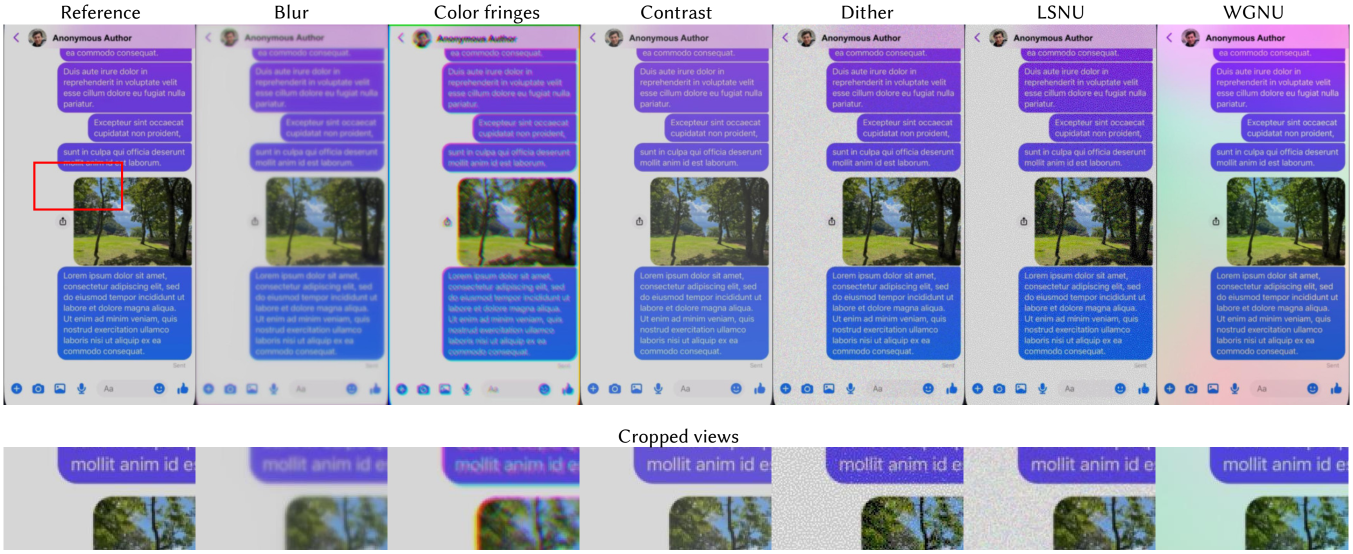

Fig. 4. *Seven types of artifacts introduced in the dataset.*

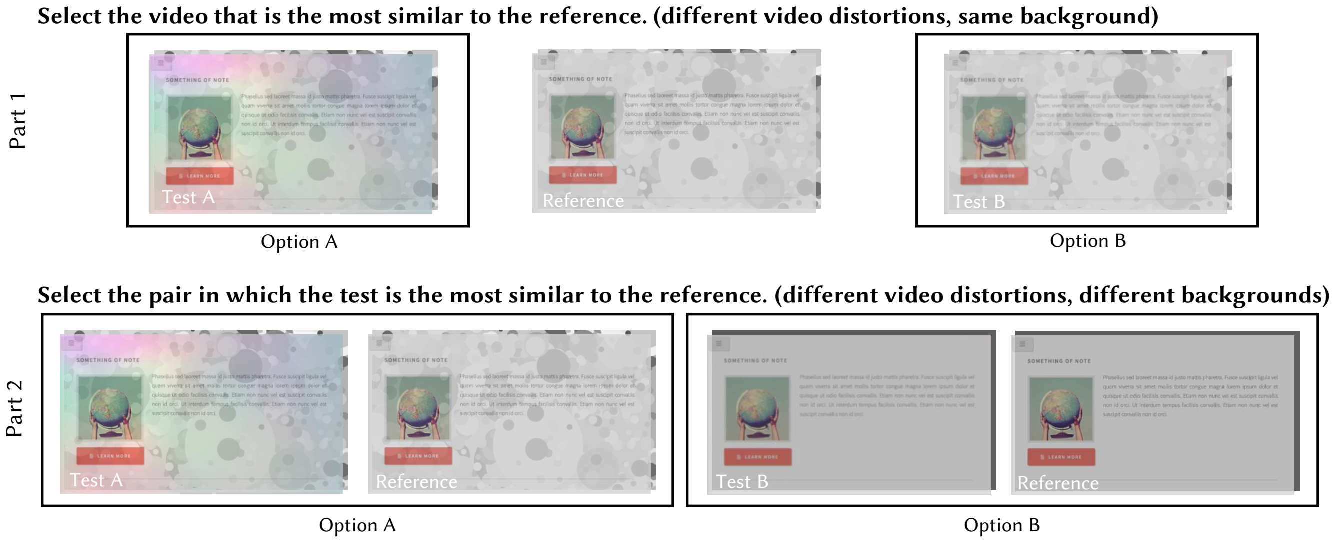
Select the video that is the most similar to the reference. (different video distortions, same background)
Part 1
Test A
Reference
Test B
Option A
Option B
Select the pair in which the test is the most similar to the reference. (different video distortions, different backgrounds)
Part 2
Test A
Reference
Test B
Reference
Option A
Option B

Fig. 5. *Experimental procedure.* A representation of the first experiment (top row) and second experiment (bottom row), detailed in Section 3.3.

Table 1. Experimental conditions of AR-DAVID dataset.

Virtual Content	#	Background Content	#
Base videos	6	Patterns	3
Distortions	6	Luminances	2
Magnitudes	2		

We use this pattern to represent natural scenes with a scale-invariant fractal nature, like trees or soil.

- An alternative model for natural scenes is the dead leaves pattern [Lee et al. 2001]. Our image follows the implementation of Gousseau and Roueff [2003], which includes hard edges between overlapping circles. This stimulus represents scenes with large contrast edges, such as occlusions.

Table 1 summarizes all experimental conditions.

3.3 Experimental Procedure

We employed a 2 interval-foced-choice (2IFC) pairwise comparison protocol with reference [Perez-Ortiz and Mantiuk 2017] to measure the visible degradation in quality due to distortions. ASAP active sampling [Mikhailiuk et al. 2021] was used to reduce the number of comparisons by scheduling the pairs of conditions that resulted in the largest information gain based on previous results.

Part 1. The first study involved comparing conditions presented on the same background. This easier experimental task yields the best accuracy for within-background comparisons. The reference was always shown first, but users could navigate between the two test videos and the reference at will afterwards.

Part 2. This study aimed to refine consistency cross-backgrounds. Two test-reference pairs are presented, containing the same background within-pair. The backgrounds between pairs could differ arbitrarily (pattern and luminance). The user is tasked with answering which pair is more alike. Users always saw the reference for each pair first, and were then able to toggle to the test video.

The experiment procedure is shown in Fig. 5. In both experiments, users were not permitted to make a selection until they viewed all test and reference videos, or before watching for at least 5 seconds. In both parts, the base video on the virtual plane was always the same within a given condition.

Participants and Procedures. 8 participants took part in a study pilot, and an additional 55 participants joined for the main study (31 for the first part, and 24 for the second). All participants signed informed consent forms, and the experiment was approved by an independent review board. The demographics of the participants were balanced in terms of age and gender, but this information was not recorded due to the IRB privacy policies. The experimental sessions lasted an average of 50.2 minutes. Participants were screened for normal or corrected-to-normal vision, and had to pass an Ishihara color vision test. Prior to the study, users were instructed to select the video with higher quality and/or fewer distortions; i.e. the one that most resembles the reference. Following completion, a qualitative 3-question survey was collected: “How easy or difficult did you find the study? Why?”, “What was your strategy when picking a video?”, and “Other comments?”, administered orally by the study organizer. These qualitative responses were leveraged to adjust the execution of the experiment, such as introducing a chair and chin rest adjustment prior to each session in the main study.

3.4 Results

The results were scaled in a unified cross-artifact perceptual scale represented with just-objectionable-difference units (JODs) using

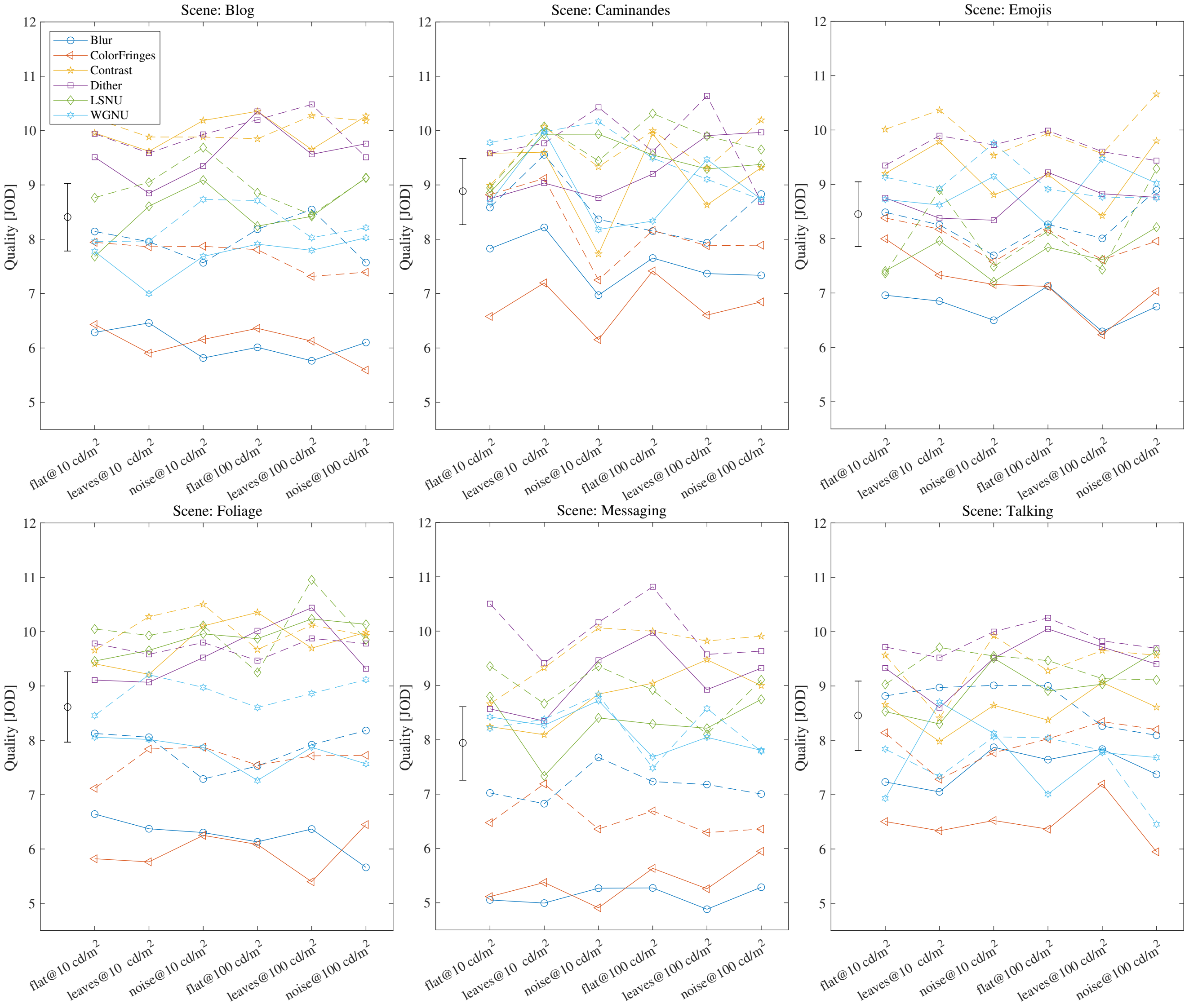

Fig. 6. *AR-DAVID results* Each line corresponds to one condition that is shown on different backgrounds. The low distortion levels are shown as dashed lines, and high distortion levels as continuous lines. The black error bar on the left of each plot denotes the average 95% confidence intervals across the conditions (not shown per condition to avoid clutter).

the *pwcmp* software[2]. The reference's quality was set to a value of 10 by convention, with lower values representing lower quality.

The results for each scene are shown in Fig. 6. The results show a substantial effect of distortion level (dashed vs. continuous lines) and per scene differences. For example, "Blur" and "Color fringes" affected the most the two scenes containing text — "Blog" and "Messaging". However, the effect of backgrounds was moderate and inconsistent across the scenes and distortions.

Our hypothesis was that high luminance backgrounds (100 cd/m^2) and those containing high-frequency patterns (noise and leaves) should mask the foreground virtual content, and therefore reduce the visibility of artifacts. The marginal distributions of the scaled JOD scores across the six backgrounds, shown in Fig. 7, indicate this effect is much smaller than expected. 2-way analysis of variance (ANOVA) showed that neither luminance of the background ($F(1, 428) = 0.12, p = 0.73$), nor the background pattern ($F(2, 428) = 0.03, p = 0.97$) resulted in significant differences in quality. This is an unexpected finding, as background luminance often exceeding the foreground should have a strong masking effect, making distortions less visible and in turn improving effective image quality scores.

Comparison with XR-DAVID. It is interesting to analyze how the video quality measured in AR differs from that measured on a regular

[2] pwcmp software: https://github.com/mantiuk/pwcmp

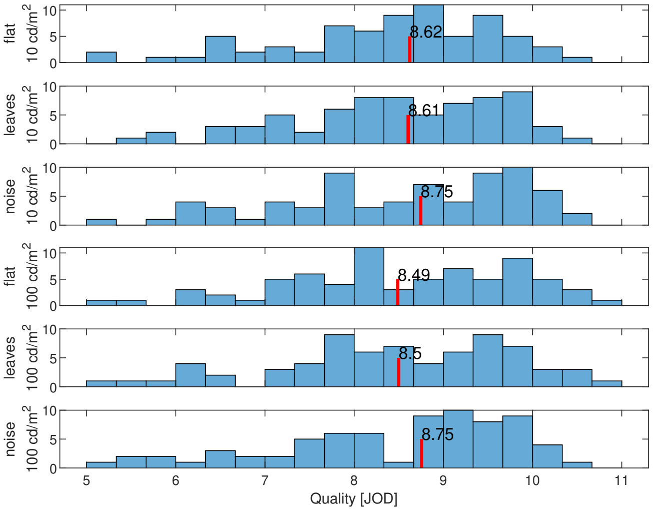

Fig. 7. *Marginal distributions* JOD scores across three background patterns and two luminance levels used in the experiment. Values marked with a red vertical line denote the median. Note that some JOD values are greater than 10 (reference) due to measurement noise.

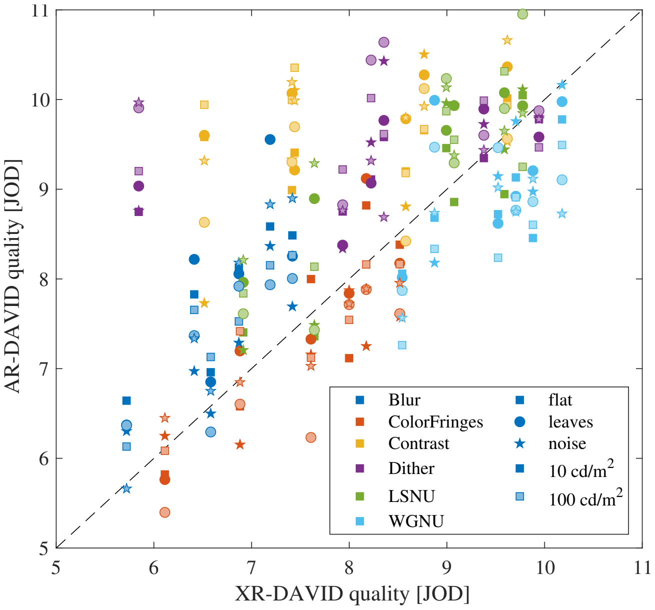

Fig. 8. *Comparison with the XR-DAVID dataset* The quality scores are compared for the three scenes that are common to both datasets. Note that for each XR-DAVID condition, there are six AR-DAVID conditions, corresponding to the six backgrounds (right column in the legend).

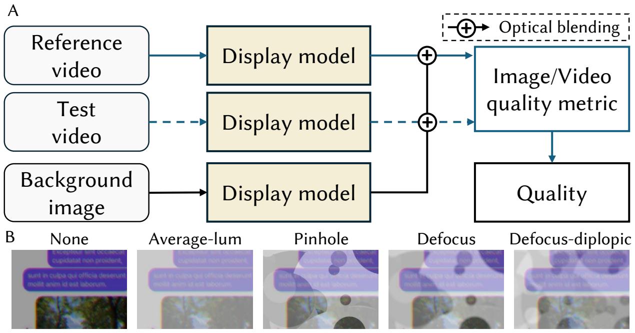

Fig. 9. *Optical blending* (A) Illustration that demonstrates the background fusion into the full-reference video quality evaluation pipeline. (B) Five of the six different optical blending schemes (none, mean, pinhole, defocus, and defocus-diplopic) are applied to the test image of *Messaging*, which has a *color fringes* artifact, and the background image of bright leaves.

display. We can achieve this by comparing our results with the XR-DAVID dataset [Mantiuk et al. 2024], as three of the video clips used (including base video, distortion types, and levels) were identical.

The scatter plot of quality values from both datasets, shown in Fig. 8, indicates that two distortions, namely *Contrast* and *Dither*, received consistently higher quality scores in AR-DAVID. This suggests that these two distortions are less noticeable when seen on an OST-AR display. This is an expected outcome as both distortions are the most noticeable in the dark portion of the image, which was the most affected by the background light. *Contrast* elevates the black level, and *Dither* was introduced in the linear space, making it more noticeable for darker tones. The other distortion types are only moderately affected by the background in an OST-AR display. We also do not observe a consistent trend for different background patterns and their luminance levels, as indicated by the ANOVA results above.

4 EVALUATION OF QUALITY METRICS

Although many image and video quality metrics exist, none of them were developed to model the perception of AR multi-focal scenes. Therefore, in this section, we explore how well the existing image and video quality metrics can predict the AR-DAVID dataset. To adapt existing metrics to AR content, we use an evaluation pipeline (Fig. 9(A)) that models the OST-AR content as seen on an AR display (Fig. 9(B)). We compared six different approaches:

- none — The metrics operate exclusively on the foreground content, and the background content information is discarded.
- average-lum — The background content is approximated by a uniform field with luminance equal to the average of the actual background. This approach simulates scenarios where a detailed representation of the background cannot be obtained (e.g. due to excessive power costs of continuously running a camera), but its luminance can be measured with a less costly ambient light sensor.
- pinhole — Foreground and background are added together in a colorimetric linear (RGB) color space, assuming pinhole optics (no defocus blur).
- pinhole-diplopic — The same as above, but the background is seen as a double image. This assumes it cannot be binocularly fused (e.g. due to effective disparity being too large).

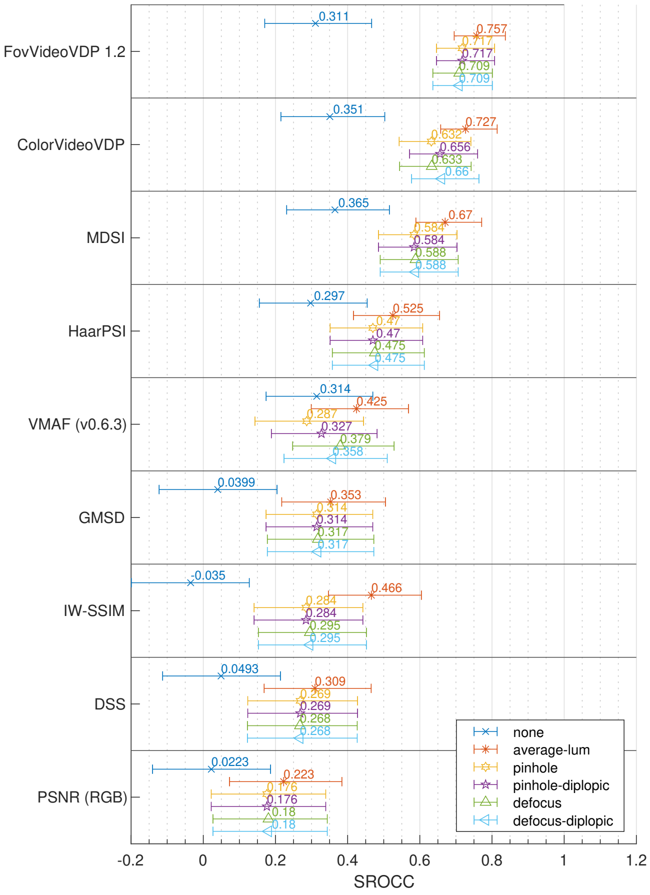

Fig. 10. *Performance of the image/video quality metrics on our AR-DAVID dataset,* shown as Spearman's rank-order correlation coefficient (SROCC). Different colors/markers denote different strategies of multi-focal fusion.

- defocus — Assuming that the observer is accommodated on the foreground plane and the background is affected by the resulting defocus blur.
- defocus-diplopic — A combination of the defocus and diplopic conditions.

The details of the optical blending methods are explained in the Appendix, and the equations are summarized in Table 3.

4.1 Metric results

We selected 15 state-of-the-art image and video metrics and evaluated their performance on the AR-DAVID dataset. The results for the 9 best-performing metrics (Table 2) are shown in Fig. 10, and detailed analysis for all 16 metrics can be found in the supplementary HTML report. For each metric, we tested the 6 optical blending methods discussed in Section 4. The metrics that can handle linear color values as output by the blending methods, such as ColorVideoVDP or FovVideoVDP, were used directly. For all other metrics, expecting display-encoded (gamma-encoded) color values, we used the PU21-encoding [Mantiuk and Azimi 2021] to ensure an appropriate range of values and perceptual uniformity.

Table 2. Quality metrics used in our tests (8 best performing). The photometric column indicates that the metric can accept photometric (absolute linear) color values.

Metric	Photometric
ColorVideoVDP [Mantiuk et al. 2024]	Yes
DSS [Balanov et al. 2015]	No
FovVideoVDP [Mantiuk et al. 2021]	Yes
GMSD [Xue et al. 2014]	No
HaarPSI [Reisenhofer et al. 2018]	No
IW-SSIM [Wang and Li 2011]	No
MDSI [Ziaei Nafchi et al. 2016]	No
VMAF [Li et al. 2016]	No

As expected, predicting AR-DAVID proved a challenging task for most of the existing metrics. Notably, we report only Spearman's rank-order correlation coefficients (SROCC) because the correlation for most metrics was low, making the standard procedure of fitting a logistic function (needed to compute RMSE and PLCC) unstable.

We introduced optical blending as a pre-processing step to adapt existing metrics to multi-focal AR content. In Fig. 10, we can see that the correlation values improved substantially for most metrics after introducing any kind of optical blending (other than none). Surprisingly, the simple average-lum blending resulted in the highest correlation coefficients. More physically accurate optical blending methods, including defocus-diploptic, performed worse than average-lum and also did not significantly improve results when compared to a simple pinhole blending. A detailed analysis of the data (found in supplementary HTML report) revealed that the main improvement of average-lum over other methods was in conditions with the leaves background. The dark discs in this background provide "tunnels" that improve the visibility of selected artifacts, such as the elevated black level for *contrast* distortion. The subjective results suggest that most observers did not use these features when judging quality and, therefore, the simplified average-lum blending that ignored them reflected subjective data better.

Metric performance varied widely. The best performance was observed for the two metrics operating on photometric units and based on psychophysical models of human vision — ColorVideoVDP and FovVideoVDP (SROCC of 0.73 and 0.76, respectively). The good performance by ColorVideoVDP is not surprising, as this metric was calibrated to predict display distortions similar to those found in AR-DAVID. Nonetheless, it was not originally designed for OST-AR content, and the SROCC correlation under the best-performing blending model for AR-DAVID is a significant decrease from the XR-DAVID dataset [Mantiuk et al. 2024], for which the authors report an SROCC value of 0.891. Some of that decrease can be explained by the differences in the quality scores between the datasets, shown in Fig. 8. The correlation values for all non-photometric metrics are even lower, below 0.5 for all but two cases.

The reason for the metrics' poor performance can be better understood by inspecting the scatter plot of ColorVideoVDP predictions in Fig. 11-left (similar results can be observed for FovVideoVDP). First, we can observe that ColorVideoVDP predicts quality that is overall higher than the actual subjective scores for the conditions shown

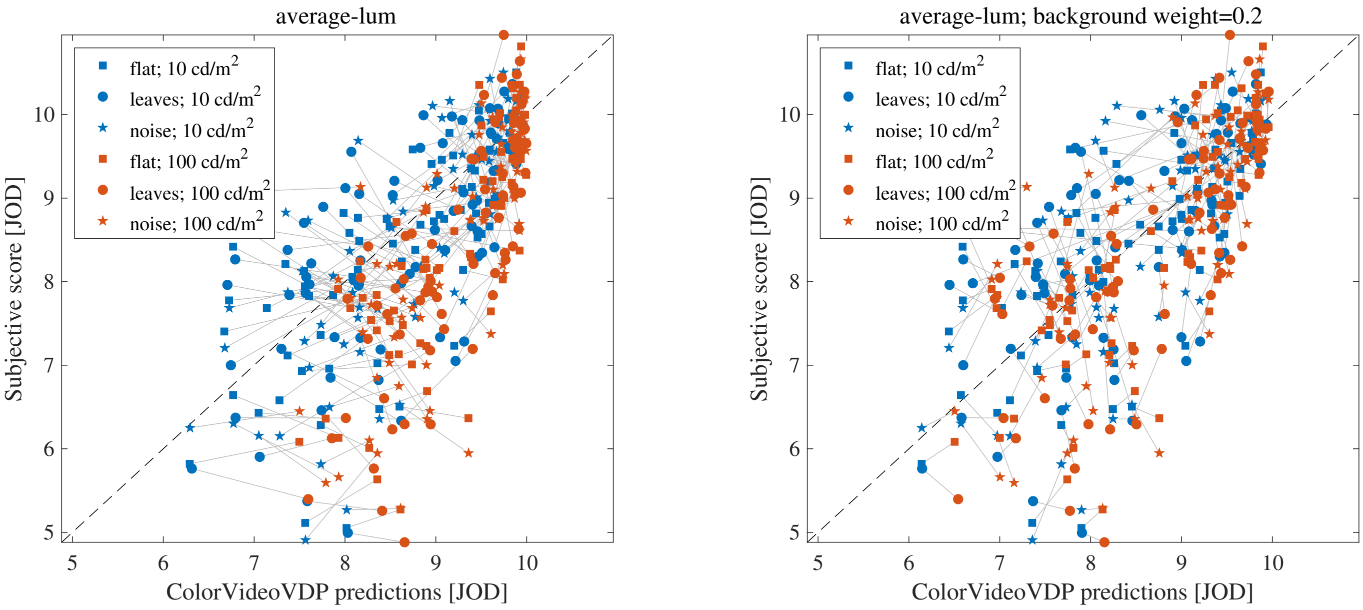

Fig. 11. *The predictions of ColorVideoVDP for* average-lum *blending (left) and for the same blending but with the background discounted to 20% (right).* The gray lines connect conditions that differ only in background luminance. Different background patterns use different marker shapes and different luminance levels use different colors.

with a bright 100 cd/m^2 background (i.e. red markers are shifted towards the right side of the plot). This means that ColorVideoVDP predicts a stronger masking effect of the background than what was found in the subjective experiment. This could indicate that users can use available cues (disparity, accommodation, and motion parallax) and disassociate the foreground from the background on an AR display. This creates an effective advantage in noticing details or distortions in the virtual plane, as compared to what would be expected in a single blended image.

Based on these results, we can conclude that strategies comprised only of optical blending can help improve metric performance but cannot fully explain the perception of content on AR displays.

Discounting background. The process of discounting the background by observers in AR is not fully understood, but some researchers suggested that it can be modeled by attenuating an image color by a scaling factor [Hassani 2019; Murdoch 2020]. We calculated FovVideoVDP and ColorVideoVDP scores for the average-lum blending and the range of background weights from 0 (equivalent to none) to 1 (equivalent to average-lum). The values of SROCC in Fig. 12-left indicate that discounting background reduces metric performance. However, the results for RMSE in Fig. 12-right indicate that performance slightly improves with the background weight of 0.3 for FovVideoVDP and 0.2 for ColorVideoVDP. Because of other factors that contribute to the uncertainty of those performance measures (error bars in Fig. 12), we cannot confirm whether this simple discounting strategy will always result in better metric performance. However, when we investigate the scatter plot in Fig. 11-right, we can see that the bias of 100 cd/m^2 background has disappeared after discounting the background with the weight of 0.2.

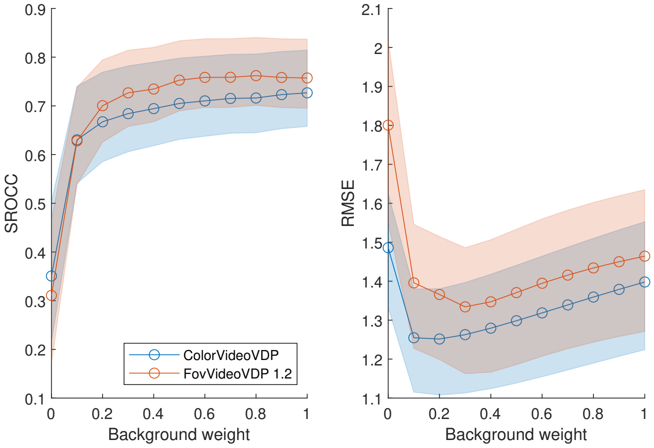

Fig. 12. *Performance of ColorVideoVDP and FovVideoVDP when the background image is discounted.* The background image is attenuated by the background weight, shown on the x-axis. The shaded region denotes 95% confidence interval.

5 CONCLUSIONS

The primary goal of this work was to obtain data that can help evaluate and develop quality metrics suitable for AR applications. To achieve this, we collected the first dataset of subjective responses to a range of AR-specific display distortions, measured using an OST-AR display using a variety of representative backgrounds. The large number of conditions (432) and highly sensitive experimental protocol using pairwise comparisons resulted in well-scaled data.

A secondary goal was to extend existing quality metrics to handle AR content with varying backgrounds. We tested several optical blending approaches, which resulted in significantly improved predictions (as compared to a baseline ignoring the background). Despite this improvement, the accuracy of existing metrics for AR content is significantly reduced, showing that there is much room for improvement.

We conclude that the perceptual models and paradigms used for regular displays may not translate directly to OST-AR displays. In particular, the perceptual aspects of the superposition of the virtual image on the background as seen through the device cannot be fully modelled as an optical mixture of light. The visual system has been shown to be capable of separating the perceived scene into "layers" to make near-accurate judgements on lightness [Gilchrist and Jacobsen 1983], brightness [Murdoch 2020], transparency [Singh and Anderson 2002], and illumination [Khang and Zaidi 2004]. Our work shows that the visual system can also partially discount the effect of the background when judging the quality of AR content, reducing the masking effect of the background. We hope this work inspires and facilitates further research on quality metrics for AR applications.

ACKNOWLEDGMENTS

We thank Cole Williams, Cameron Wood and Alex Gherman for their help in conducting the user study, as well as John Hill for his help organizing it. Additionally, we thank Curtis Torrey and Xin Li for their efforts in building and calibrating the haploscope.

REFERENCES

Tunç O. Aydın, Rafal Mantiuk, and Hans-Peter Seidel. 2008. Extending quality metrics to full luminance range images, Bernice E. Rogowitz and Thrasyvoulos N. Pappas (Eds.). San Jose, CA, 68060B. https://doi.org/10.1117/12.765095

Amnon Balanov, Arik Schwartz, Yair Moshe, and Nimrod Peleg. 2015. Image quality assessment based on DCT subband similarity. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, Quebec City, QC, Canada, 2105–2109. https://doi.org/10.1109/ICIP.2015.7351172

Scott J Daly. 1992. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, Vol. 1666. SPIE, 2–15.

Jian Ding and Dennis M Levi. 2017. Binocular combination of luminance profiles. *Journal of vision* 17, 13 (2017), 4–4.

Jian Ding and George Sperling. 2006. A gain-control theory of binocular combination. *Proceedings of the National Academy of Sciences* 103, 4 (2006), 1141–1146.

Siavash Eftekharifar, Anne Thaler, Adam O. Bebko, and Nikolaus F. Troje. 2021. The role of binocular disparity and active motion parallax in cybersickness. *Experimental Brain Research* 239, 8 (Aug. 2021), 2649–2660. https://doi.org/10.1007/s00221-021-06124-6

Chunyu Gao, Yuxiang Lin, and Hong Hua. 2012. Occlusion capable optical see-through head-mounted display using freeform optics. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 281–282.

Wilson S Geisler. 2008. Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59 (2008), 167–192.

Alan L. Gilchrist and Alan Jacobsen. 1983. Lightness constancy through a veiling luminance. *Journal of Experimental Psychology: Human Perception and Performance* 9, 6 (1983), 936–944. https://doi.org/10.1037/0096-1523.9.6.936

Yann Gousseau and François Roueff. 2003. The dead leaves model: general results and limits at small scales. *arXiv preprint math/0312035* (2003).

Nargess Hassani. 2019. *Modeling color appearance in augmented reality.* PhD dissertation. Rochester Institute of Technology.

Juan David Hincapié-Ramos, Levko Ivanchuk, Srikanth K Sridharan, and Pourang P Irani. 2015. SmartColor: real-time color and contrast correction for optical see-through head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics* 21, 12 (2015), 1336–1348.

Byung-Geun Khang and Qasim Zaidi. 2004. Illuminant color perception of spectrally filtered spotlights. *Journal of Vision* 4, 9 (Aug. 2004), 2. https://doi.org/10.1167/4.9.2

Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. 2020. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)* 39, 2 (2020), 1–12.

George-Alex Koulieris, Bee Bui, Martin S. Banks, and George Drettakis. 2017. Accommodation and comfort in head-mounted displays. *ACM Trans. Graph.* 36, 4, Article 87 (jul 2017), 11 pages. https://doi.org/10.1145/3072959.3073622

Ann B Lee, David Mumford, and Jinggang Huang. 2001. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision* 41 (2001), 35–59.

Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. 2016. *Toward A Practical Perceptual Video Quality Metric.* Technical Report. The NETFLIX Tech Blog. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

Jingyu Liu, Akshay Jindal, Claire Mantel, Soren Forchhammer, and Rafal K. Mantiuk. 2022. How bright should a virtual object be to appear opaque in optical see-through AR?. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 676–684. https://doi.org/10.1109/ISMAR55827.2022.00085

Taoran Lu, Fangjun Pu, Peng Yin, Tao Chen, Walt Husak, Jaclyn Pytlarz, Robin Atkins, Jan Frohlich, and Guan-MingSu. 2016. ITP Colour Space and Its Compression Performance for High Dynamic Range and Wide Colour Gamut Video Distribution. *ZTE Communications* 1, 1 (2016), 1–7.

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 Transactions on graphics (TOG)* 30, 4 (2011), 1–14.

Rafal 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)*. IEEE, 1–5. https://doi.org/10.1109/PCS50896.2021.9477471

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 Transactions on Graphics (TOG)* 40, 4 (2021), 1–19.

Rafał K Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. 2024. ColorVideoVDP: A visual difference predictor for image, video and display distortions. *ACM Transaction on Graphics* 43, 4 (2024), 129. https://doi.org/10.1145/3658144

Nathan Matsuda, Alex Chapiro, Yang Zhao, Clinton Smith, Romain Bachy, and Douglas Lanman. 2022. Realistic luminance in vr. In *SIGGRAPH Asia 2022 Conference Papers*. 1–8.

Christoffer Menk and Reinhard Koch. 2012. Truthful color reproduction in spatial augmented reality applications. *IEEE Transactions on Visualization and Computer Graphics* 19, 2 (2012), 236–248.

Fabio Metelli. 1974. The perception of transparency. *Scientific American* 230, 4 (1974), 90–99.

Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K Mantiuk. 2021. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2559–2566.

Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. 2024. Perceptual Video Quality Assessment: A Survey. *arXiv preprint arXiv:2402-03413* (2024).

Michael J Murdoch. 2020. Brightness matching in optical see-through augmented reality. *JOSA A* 37, 12 (2020), 1927–1936.

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 (July 2015), 46–60. https://doi.org/10.1016/j.image.2015.04.009 Citation Key: Narwaria2015.

Yana Nehmé, Johanna Delanoy, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. 2023. Textured Mesh Quality Assessment: Large-scale Dataset and Deep Learning-based Quality Metric. *ACM Transactions on Graphics* 42, 3 (June 2023), 1–20. https://doi.org/10.1145/3592786

Maria Perez-Ortiz and Rafal K Mantiuk. 2017. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712-03686* (2017).

Kishore Rathinavel, Gordon Wetzstein, and Henry Fuchs. 2019. Varifocal occlusion-capable optical see-through augmented reality display based on focus-tunable optics. *IEEE transactions on visualization and computer graphics* 25, 11 (2019), 3125–3134.

Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. 2018. A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication* 61 (Feb. 2018), 33–43. https://doi.org/10.1016/j.image.2017.11.001

Daniel Ruderman and William Bialek. 1993. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems* 6 (1993).

Stephen Sebastian, Johannes Burge, and Wilson S Geisler. 2015. Defocus blur discrimination in natural images with natural optics. *Journal of Vision* 15, 5 (2015), 16–16.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. 2005. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application* 30, 1 (2005), 21–30.

Manish Singh and Barton L. Anderson. 2002. Toward a perceptual theory of transparency. *Psychological Review* 109, 3 (2002), 492–519. https://doi.org/10.1037/0033-295X.109.3.492

Larry N Thibos, Ming Ye, Xiaoxiao Zhang, and Arthur Bradley. 1992. The chromatic eye: a new reduced-eye model of ocular chromatic aberration in humans. *Applied optics* 31, 19 (1992), 3594–3600.

David J Tolhurst, Yoav Tadmor, and Tang Chao. 1992. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics* 12, 2 (1992), 229–232.

Minqi Wang, Jian Ding, Dennis M. Levi, and Emily A. Cooper. 2024. The Effect of Interocular Contrast Differences on the Appearance of Augmented Reality Imagery. *ACM Transactions on Applied Perception* 21, 1 (Jan. 2024), 1–23. https://doi.org/10.1145/3617684

Zhou Wang and Qiang Li. 2011. Information Content Weighting for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 5 (May 2011), 1185–1198. https://doi.org/10.1109/TIP.2010.2092435

Andrew B Watson and John I Yellott. 2012. A unified formula for light-adapted pupil size. *Journal of vision* 12, 10 (2012), 12–12.

Krzysztof Wolski, Laura Trutoiu, Zhao Dong, Zhengyang Shen, Kevin Mackenzie, and Alexandre Chapiro. 2022. Geo-Metric: A Perceptual Dataset of Distortions on Faces. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–13.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. 2014. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Transactions on Image Processing* 23, 2 (Feb. 2014), 684–695. https://doi.org/10.1109/TIP.2013.2293423

Lili Zhang. 2022. *Lightness, Brightness, and Transparency in Optical See-Through Augmented Reality.* PhD dissertation. Rocherster Institute of Technology.

Yunjin Zhang, Rui Wang, Yifan Peng, Wei Hua, and Hujun Bao. 2021. Color contrast enhanced rendering for optical see-through head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 4490–4502.

Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. 2016. Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator. *IEEE Access* 4 (2016), 5579–5590. https://doi.org/10.1109/ACCESS.2016.2604042

Junyu Zou, Qian Yang, En-Lin Hsiang, Haruki Ooishi, Zhuo Yang, Kifumi Yoshidaya, and Shin-Tson Wu. 2021. Fast-response liquid crystal for spatial light modulator and LiDAR applications. *Crystals* 11, 2 (2021), 93.

APPENDIX: OPTICAL BLENDING

The experimental setup is illustrated in Fig. 1, where two planes represent the foreground (FG) and the background (BG). We assume both planes are orthogonal to the visual axis and aligned. Each image is converted to the CIE 1931 XYZ color space to ensure linearity, based on modeling the display used in the experiment (see [Mantiuk et al. 2024, Eq. 2]). The XYZ tristimulus values of the perceived image (C_{eff}) are then calculated as the sum of the values for foreground (FG) and background (BG)each plane (C_{FG} and C_{BG}):

$$C_{\text{eff}} = C_{\text{FG}} + C_{\text{BG}}. \tag{1}$$

Here, the C_{eff}, C_{FG}, and C_{BG} represent two-dimensional functions of visual angle per color channel, but the notations of visual angle are omitted here for brevity.

Focal fusion. The stimuli in our experiment consist of two planes, each at different distance from the eye. The near (foreground) plane is at the diopter distance of D_0 and the far (background) plane is at the distance $D_0 + \Delta D$ (distances shown in Fig. 2). Because the plane separation (0.73 D) is greater than the blur discrimination threshold (between 0.125 D and 0.625 D, depending on the frequency spectrum of the images [Sebastian et al. 2015]), we can assume users focusing on the foreground would perceive the background image as blurred.

The fused image (XYZ color space) with the defocus blur is computed as:

$$C_{\text{eff}} = C_{\text{FG}} + C_{\text{BG}} * h, \tag{2}$$

where, h represents the 2D point spread function (PSF), and $*$ represents a 2D convolution operator. The PSF is presented in a domain of visual angle ($\vec{u}$) in radian units, with parameters of the dioptric difference ΔD, and pupil diameter of p (in metric units), as follows:

$$h(\vec{u}; \Delta D, p) = \begin{cases} 1, & \|\vec{u}\| < p\Delta D/2, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The PSF is estimated based on the mean luminance and size of the stimuli, using the model of Watson and Yellott [2012]. $\|\cdot\|$ denotes the l-2 norm of the given vector. Here, the PSF is estimated without accounting for human eye aberrations [Thibos et al. 1992] or diffraction effects from the pupil aperture, as the images are blended in the display plane rather than the retinal plane. Note that the quality metrics operate on the displayed image.

Table 3. Summary of optical blending models. The $mean()$ operation on the tristimulus values represents a channel-dependent averaging operation.

Blending model	Equation
none	$C_{\text{eff}} = C_{\text{FG}}$
average-lum	$C_{\text{eff}} = C_{\text{FG}} + mean\,(C_{\text{BG}})$
pinhole	$C_{\text{eff}} = C_{\text{FG}} + C_{\text{BG}}$
defocus	$C_{\text{eff}} = C_{\text{FG}} + C_{\text{BG}} * h$
pinhole-diplopic	$C_{\text{eff}} = C_{\text{FG}} + \frac{1}{2} \sum_{k \in \{\text{L,R}\}} C_{\text{BG}}(\vec{u} - \vec{u}_k)$
defocus-diplopic	$C_{\text{eff}} = C_{\text{FG}} + \frac{1}{2} \sum_{k \in \{\text{L,R}\}} C_{\text{BG}}(\vec{u} - \vec{u}_k) * h(\vec{u}; \Delta D, p)$

Binocular fusion. The disparity arising from the distance between the foreground and background and the observer's interpupillary distance (IPD) may cause the background to be perceived as a double (diplopic) image. The image perceived by each eye ($C_{\text{L/R}}$: XYZ tristimulus values as perceived by left or right eye) can be simulated based on the geometry of the display system, assuming gaze location at the center point of the foreground display. Since there is no disparity between images positioned at the virtual plane's depth, only the background image is shifted for each eye, and the corresponding values can be averaged as follows:

$$C_{\text{BG}} = \frac{1}{2} \sum_{k \in \{\text{L,R}\}} C_{\text{BG}}(\vec{u} - \vec{u}_k), \tag{4}$$

with the angular shift of the visual axis of the left or right eye given as $\vec{u}_{\text{L/R}} = \left(\tan^{-1}(\mp \frac{ipd}{2} D_0), 0\right)$. Although the IPD can vary individually and effectively change with gaze direction [Konrad et al. 2020], we make a simplifying assumption and set the effective IPD to a representative value of 63 mm [Khang and Zaidi 2004]. In lieu of sophisticated models for binocular fusion [Ding and Levi 2017; Ding and Sperling 2006], we employ a simple linear summation model of binocular fusion in a linear color space. The formulations of all optical blending strategies used in this work are provided in Table 3.