Journal of Perceptual Imaging® 7: 1–12, 2024.

Adjusting Transparency Toward Optimizing Face Appearance in Optical See-Through Augmented Reality

Sofie R. Herbeck, Michael J. Murdoch, Christopher A. Thorstenson

Munsell Color Science Laboratory, Rochester Institute of Technology, Rochester, NY 14623, United States
E-mail: srh4649@rit.edu

Abstract. *Augmented reality (AR) combines elements of the real world with additional virtual content, creating a blended viewing environment. Optical see-through AR (OST-AR) accomplishes this by using a transparent beam splitter to overlay virtual elements over a user's view of the real world. However, the inherent see-through nature of OST-AR carries challenges for color appearance, especially around the appearance of darker and less chromatic objects. When displaying human faces—a promising application of AR technology—these challenges disproportionately affect darker skin tones, making them appear more transparent than lighter skin tones. Still, some transparency in the rendered object may not be entirely negative; people's evaluations of transparency when interacting with other humans in AR-mediated modalities are not yet fully understood. In this work, two psychophysical experiments were conducted to assess how people evaluate OST-AR transparency across several characteristics including different skin tones, object types, lighting conditions, and display types. The results provide a scale of perceived transparency allowing comparisons to transparency for conventional emissive displays. The results also demonstrate how AR transparency impacts perceptions of object preference and fit within the environment. These results reveal several areas with need for further attention, particularly regarding darker skin tones, lighter ambient lighting, and displaying human faces more generally. This work may be useful in guiding the development of OST-AR technology, and emphasizes the importance of AR design goals, perception of human faces, and optimizing visual appearance in extended reality systems.*
Keywords: *augmented reality, optical see-through AR, transparency, perception, faces*

[DOI: 10.2352/J.Percept.Imaging.2024.7.000404]

1. INTRODUCTION

Augmented reality (AR) aims to combine visual elements of the real world with additional virtual content, allowing users to perceive some aspect of the physical world that is "augmented" by virtual elements. There are currently two major distinct optical paradigms that approach this goal: optical see-through AR (OST-AR), and video see-through AR (VST-AR). In OST-AR, the physical world is viewed directly through a transparent pane (i.e., a beam splitter), on which additional digital objects can be displayed, with the goal of making them appear as situated within the physical space (see Figure 1). In VST-AR the physical world is viewed indirectly, by showing a real-time video feed on a near-eye emissive display, within which additional digital content can be superimposed. Each approach has its tradeoffs; VST-AR's fully-emissive display can seamlessly integrate the AR component with the live video, but the indirect method of viewing the physical environment may reduce immersion, and presents issues with parallax, vergence, and "cybersickness," which can include disorientation, headaches, and nausea [4, 9]. OST-AR allows users to more directly interact with the physical world, and alleviates the cybersickness concern, but introduces additional challenges for displaying the augmented digital content. Of significance among these are color reproduction and transparency. Because the transparent optics rely on an additive light model (additional light is reflected by a transparent pane and added to the mixed environment), it is particularly difficult to display augmented content having darker and less chromatic colors. This is because darker objects add much less light the system, which may be insufficient when combined with the ambient light of the physical environment. The result is that augmented content struggles to achieve full opacity, with darker objects appearing appreciably more transparent [29]. This challenge inherent to OST-AR display is the focus of the present work.

Virtual social interaction is a promising application for AR technology, with uses for teleconferencing, social meeting spaces, and medical interactions. A crucial aspect of this application will involve reliable AR reproduction of human faces. Faces represent a socially relevant class of stimuli that have been a major topic in perception research [11]. People regularly and frequently extract important social information from faces, including perceptions of age, gender, health, and emotion [10, 12, 16, 23]. Past work has shown that people are particularly sensitive to processing the color appearance of human faces [7, 26, 27], and that even small changes to facial color appearance can impact how other people evaluate them [1, 25, 28]. The OST-AR challenges previously described present a particular obstacle for reliably reproducing AR human faces. Because skin tones naturally vary considerably across the human population, there is likely to be a disparity in adequate AR display of faces that is particularly detrimental concerning faces having darker skin tones. This discrepancy can introduce an unintentional bias which is likely to limit equitable access to and experiences in OST-AR technology across the human population. Indeed, it has been determined earlier that OST-AR displays reproduce

Received May 20, 2024; accepted for publication Oct. 9, 2024; published online Oct. 23, 2024. Associate Editor: Adar Pelah.

2575-8144/2024/7/000404/12/$00.00

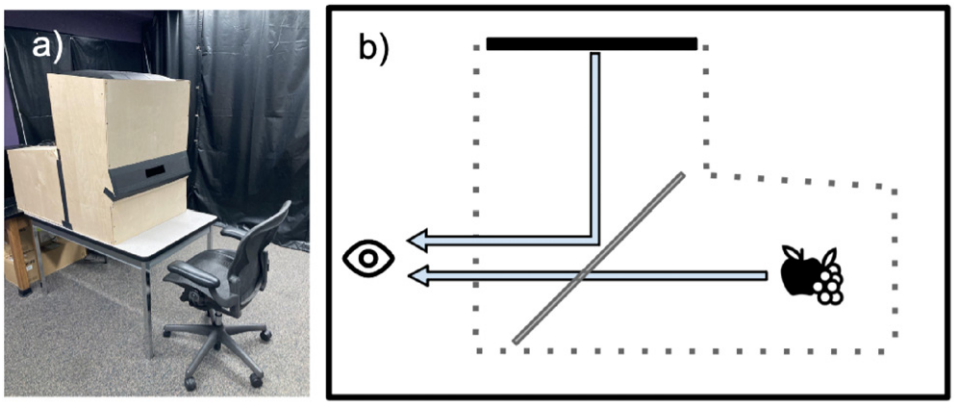

Figure 1. OST-AR setup used in this study. (a) The setup is housed in a wooden tabletop booth featuring a front viewing slit. (b) Diagram of internal arrangement; the AR content from the top display is overlaid onto physical objects in rear via slanted beam splitter.

faces with lighter skin tones more favorably than those with darker skin tones [20], and that substantial visual adjustments are needed to equitably reproduce faces with darker skin tones [2]. These differences are largely driven by the considerable transparency induced by darker OST-AR skin tones, especially when ambient lighting of the physical environment is very bright. As this technology becomes more widely-adopted, it is paramount that human faces are reproduced in an equitable and desirable way. This study focuses largely on perceptions of facial stimuli, while comparing these to non-face objects.

Previous work has aimed to assess and model perceptual transparency and brightness matching with isolated color patches or abstract objects in OST-AR contexts [19, 29]. The two primary avenues for reducing potentially-problematic transparency in OST-AR are occlusion and advanced rendering techniques. Occlusion-capable OST-AR headsets counter transparency by modulating their display's opacity to block the physical scene in the region of a rendered object, however the additional optics involved are generally bulky, complex, and expensive [5]. Advanced rendering techniques, such as real-time color and/or contrast correction, can modify rendered stimuli to make them appear more salient regardless of the display's transparency—but live rendering introduces latency issues and increases the computational demands on the headset itself, akin to those of VST-AR [8].

This study investigates the transparency perception of human faces in OST-AR, and evaluates whether transparency is inherently detrimental in all contexts. First, in Experiment 1, we assess a perceived transparency scale that allows the mapping of perceived transparency in OST-AR to that of a conventional emissive opacity parameter via alpha compositing. To induce transparency in OST-AR, we manipulate image lightness gamma, named for the exponent in the power law image transformation used to adjust image brightness, which has been previously used as a proxy parameter for transparency [15]. This perceived transparency scale is assessed among combinations of stimulus shapes (faces and non-faces), stimulus colors (varying skin tones), and background lightnesses simulating variable ambient lighting conditions (light and dark checkerboards). Additionally, we assess how perceived transparency—for both OST-AR and conventional emissive displays—impacts people's evaluation (i.e., perceived "visual acceptability") of the reproduced objects.

Then, in Experiment 2, we assess how OST-AR transparency influences evaluations relating to two distinct AR design goals: preferred appearance (i.e., how preferable the objects appear within the given environment), and environmental fit (i.e., how well the AR objects appear to be situated within the physical environment). We chose to evaluate these design goals as it is possible that transparency in OST-AR may be more detrimental to one than to the other, rather than having a uniform impact on perception across them. One motivation for this experiment comes from the premise that the "real" environment and the virtual content might often come from physical lighting conditions having different chromaticities, which may impede evaluations of the scene when they are combined. Therefore, we additionally assessed these evaluations among combinations of lighting color (cool, warm, and magenta), physically in the viewing booth and simulated for the AR stimuli.

2. EXPERIMENT 1

Experiment 1 comprised two parts, which together allowed the development of both a perceptual transparency function that maps perceptual OST-AR transparency to emissive image transparency, as well as a threshold indicating the transparency values on that function which were deemed visually acceptable. These functions and thresholds were expected to vary based on stimulus color and background lightness. We additionally explored whether they varied between face and non-face stimuli.

A tabletop OST-AR setup was used (Fig. 1), utilizing a large beam splitter to optically combine views of a 5-channel LED-illuminated viewing booth and a high-brightness LCD (referred to as the emissive—as opposed to AR—display). From the viewer's perspective, the reflected image of the emissive display appears to float transparently within the space of the viewing booth. In this experiment, the booth

Table I. Measured luminance (Y) values for lighter and darker patches on both checkerboards, with computed Michelson contrast values.

Checkerboard	Lighter Y	Darker Y	Michelson contrast
Light	150.51	132.74	0.06
Dark	11.87	10.23	0.07

contained checkerboard panels as background for the AR stimuli, as well as a second emissive display for reference. This emissive display was placed in the OST-AR viewing booth, situated to appear to be in the same 3D plane as displayed AR elements. This was used to compare observers' perception of the AR stimuli's transparency to that of a more traditional emissive display.

As discussed in previous work, transparency is difficult to quantify perceptually without some sort of contrast pattern behind the stimulus whose transparency is modulated [29]. For this experiment, achromatic checkerboard patterns were created to serve as backgrounds (see Table I).

Three faces with different skin colors were selected, with decreasing average skin lightness, from a set of AI-generated face images [6]. Skin lightness was averaged over combined forehead and cheek areas using both CIELAB L^* (using D65 at 438 cd/m^2 as the white point) and Individual Typology Angle (ITA), which corresponds to an estimation of melanin content in skin [30]. These values were used to select 3 skin tones that were linearly spaced along these dimensions. The selected skin tones had average CIELCh values as follows: light [$L = 80.46$, $C = 19.72$, $h = 56.54$], mid-tone [$L = 60.68$, $C = 37.44$, $h = 59.06$], and dark [$L = 44.22$, $C = 25.59$, $h = 53.92$]. Three matching non-face stimuli were rendered using the Glavens dataset [22]. This collection of 3D objects utilizes structured randomness to create a scale of perceptual complexity [21]. A glaven with medium complexity was selected to mimic the contours of a face while still appearing as an abstract shape. The average skin color of each of the 3 faces was computed and applied to each of the 3 glavens. All 6 stimuli are shown in Figure 2.

2.1 *Methods*

Checkerboard backgrounds were used, with a lighter and a darker version to assess differences in AR transparency perception between different background luminance levels (to simulate variable ambient lighting). The contrast level was equalized between both backgrounds using Michelson contrast [18], represented as C:

$$C = \frac{\Delta Y}{2\overline{Y}}, \tag{1}$$

where Y is luminance in cd/m^2, as measured by a CR-250 spectroradiometer. Final patch luminance values and Michelson contrast values (of approximately 0.07) are shown in Table I. A paper background was printed, measured, and subsequently matched on the color-characterized emissive display.

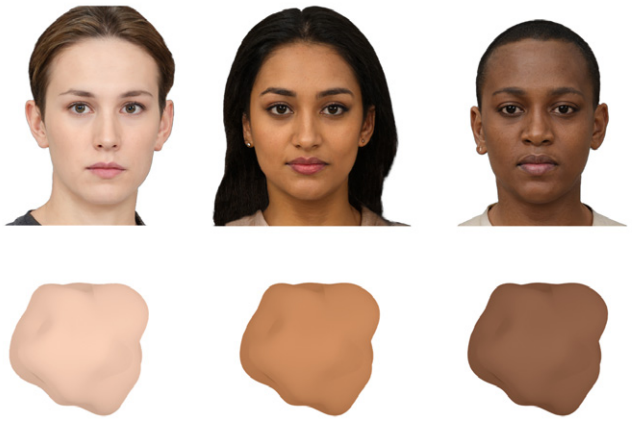

Figure 2. The stimulus set comprised faces and non-face objects having 3 matched skin tones.

The printed background and the emissive display were framed with black mat board to make them appear as similar as possible to each other. The printed background was placed such that the AR stimulus would appear to be in-plane with its background, mimicking the integrated stimulus and background shown alongside it via the emissive display. Room lighting and AR viewing booth lighting were kept constant at an approximation of D65 (consistent with the lighting in the rendered images). The experimental setup with stimuli as seen by the observer is shown in Figure 3.

Twenty five observers participated in this experiment, comprising university students and faculty, primarily within the Program of Color Science at the Rochester Institute of Technology. No observers indicated having a color vision deficiency, and all had normal or corrected-to-normal vision. Demographic data was not collected in Experiment 1.

2.1.1 *Expt 1A: Matching Perceived Transparency*

In the first part of the experiment, observers were asked to complete a method-of-adjustment psychophysical task. For each trial, they were shown one corresponding stimulus on both AR and emissive displays within the viewing booth (see Fig. 3c/d), with the emissive display's stimulus being fixed at one of 7 emissive alpha levels evenly spaced between 0.1 (so transparent as to be barely visible) and 1 (fully opaque). Observers were tasked with adjusting the gamma level of the AR stimulus until a perceptual transparency match was produced. Transparency was defined as "the amount of background contrast pattern that can be seen through the stimulus." Gamma adjustment was applied only on the L* channel of each CIELAB-converted image, to minimize the hue shift seen at extreme gamma values when adjusting gamma in RGB. While the gamma adjustment may alter color appearance due to gamut limitations, especially at the highest levels of gamma, participants were instructed to focus solely on perceived transparency matching, and to ignore color appearance matching. Stimuli were presented over both light and dark backgrounds, with two repetitions of each [stimulus color (3), stimulus shape (2), background (2), alpha

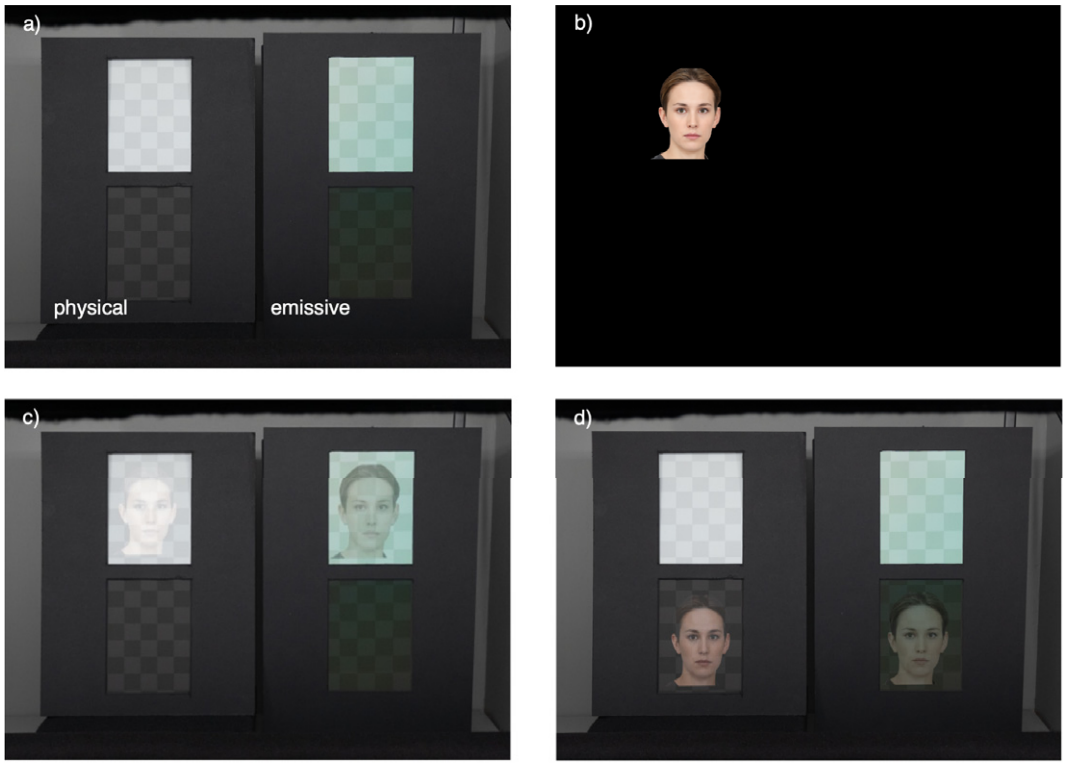

Figure 3. The experimental setup was composed of background and AR components. (a) Checkerboard backgrounds that were physically printed (left) and digitally shown on an emissive display (right), visible in transmission through the beam splitter. (b) Example of AR stimulus visible in reflection through the beam splitter. (c) Example of an observer's view with the transparent AR stimulus overlaid on the light printed background checkerboard (left) and the matching stimulus alpha-composited on the light emissive background (right). (d) Example of an observer's view of AR and emissive stimuli, but over the dark backgrounds. Note: the emissive display's color is not accurately reproduced in this image, but was confirmed to colorimetrically match the printed samples and AR stimuli when measured using a CR-250 spectroradiometer.

level (7)] combination, for a total of 168 trials. If observers reached the upper or lower bound of allowed gamma values and still felt that the AR stimulus did not perceptually match the emissive display's stimulus in terms of transparency, they were instructed to choose the alternate submission key to flag the trial as being out of gamma-gamut for transparency matching.

2.1.2 *Expt 1B: Assessing Visual Acceptability*

In the second part of the experiment, observers were shown stimuli individually on the AR or emissive display, at seven emissive alpha levels from 0.1–1 and at seven AR gamma levels from 0.2 to 3 (constrained by previous work [15]), totaling an additional 168 trials. In this section, observers were asked to judge whether each individual stimulus presentation was a "visually acceptable" representation of that stimulus in a virtual setting. The gamma range sampling allowed us to identify the threshold above which transparency matches were possible, but the stimulus no longer looked "good enough". For example, this could include cases where darker stimuli could be made to appear opaque on a given background, but only at the cost of extreme gamma boosting and hence perceptual desaturation.

The alpha sampling allowed us to assess the transparency below which stimuli were simply not sufficiently visible, in which case matched gamma values below this value should be excluded.

2.2 *Results*

2.2.1 *Expt 1A: Matching Perceived Transparency*

For subsequent analyses, the observer-adjusted parameter gamma value was transformed to a more perceptually linear and intuitive quantity. As the visual effect of gamma is an overall increase or decrease in image lightness, the resulting difference in the mid-scale lightness was computed, referred to as mid-scale ΔL^* (ΔL^* henceforth). For a given gamma value, ΔL^* is the difference between the resulting L^* and 50, for an input L^* of 50. For orientation, a gamma value of 1 is an identity function, resulting in $\Delta L^* = 0$; gamma values less than 1 result in positive ΔL^* (brighter AR stimuli), and gamma values greater than 1 result in negative ΔL^* (darker AR stimuli).

Repeated measures ANOVA was done first to evaluate how participants adjusted lightness of the AR stimuli to match the perceived transparency of emissive stimuli at 7 levels of emissive alpha (highly transparent to fully opaque), as well as how these adjustments were influenced by stimulus shape (2; faces versus glavens), stimulus color (3; light, mid-tone, dark), and background lightness (2; light versus dark). Following this, linear mixed-effects models with random by-participant slopes and intercepts accounting for

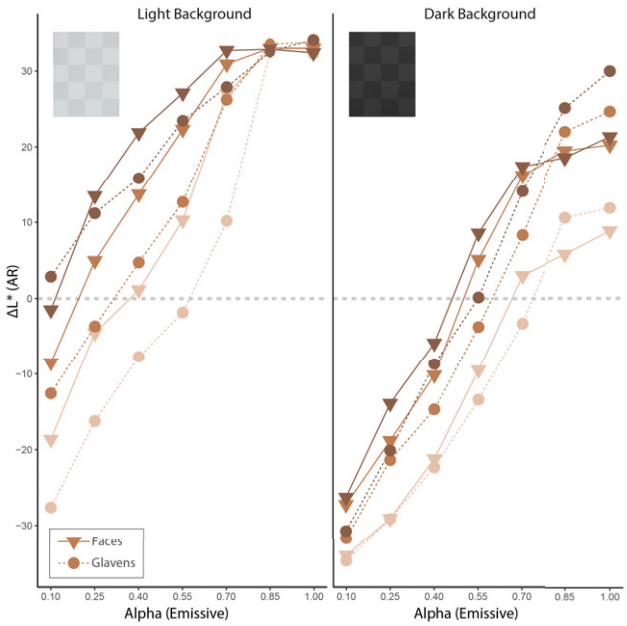

Figure 4. Perceptual transparency matches from emissive to AR display, reporting adjusted ΔL^* as a function of emissive display alpha level. Points are plotted by skin color and shapes correspond to the stimulus shape. Observers increased brightness in AR from the baseline to match many opacity values, especially for darker stimuli and as emissive opacity increased. Lighter backgrounds required more brightness boosting to make AR stimuli appear similarly opaque to emissive displays.

repeated measures were conducted to evaluate any changes among the linear effects of emissive alpha on participant adjustments as a function of the independent variables. The dependent variable was mid-scale ΔL^*. A summary of the transparency matching results is shown in Figure 4. There was a significant main effect of emissive alpha on ΔL^*, $F(6, 144) = 1016.36$, $p < 0.001$, indicating that as targets' emissive alpha increased, participants increased L^* to match their perceived transparency ($B = 55.46$, SE $= 1.41$, $p < 0.001$). There was a significant effect of stimulus shape, $F(1, 24) = 75.47$, $p < 0.001$, indicating that L^* was increased more for faces ($M = 7.57$, SE $= 0.496$) than for glavens ($M = 4.23$, SE $= 0.443$) to match the targets' perceived transparency. There was a significant effect of stimulus color, $F(2, 48) = 1175.31$, $p < 0.001$, on ΔL^* adjustments. Greater L^* increases were needed to match perceived transparency for dark stimuli ($M = 12.01$, SE $= 0.43$) than for mid-tone stimuli ($M = 7.58$, SE $= 0.52$), $t(24) = 15.44$, $p < 0.001$, than for light stimuli ($M = -1.90$, SE $= 0.43$), $t(24) = 38.37$, $p < 0.001$. There was a significant effect of background lightness, $F(1, 24) = 854.40$, $p < 0.001$, indicating that L^* needed to increase more to match targets' perceived transparency when stimuli were viewed on a light background ($M = 15.10$, SE $= 0.58$) than on a dark background ($M = -3.30$, SE $= 0.48$). However, these effects were qualified by additional significant interaction effects among the independent variables (see Table II). Upon further investigation of these interactions, we determined that the general directional patterns described above remained, but that the effect sizes differed among levels of the independent variables. This can be seen in Figure 5, which shows that L^* needed to be increased more in order to match the perceived transparency of emissive targets over the lighter versus the darker background, as well as for darker stimuli, with slightly larger increases of L^* for faces than glavens.

Table II. Summary of main effects and interactions found in Expt 1.

	Light background			Dark background		
Effect	*df*	*F*	*p*	*df*	*F*	*p*
Alpha	6, 144	526.21	<0.001	6, 144	667.34	<0.001
Shape	1, 24	63.88	<0.001	1, 24	7.17	0.013
Color	2, 48	589.06	<0.001	2, 48	699.59	<0.001
Shape*color	2, 48	31.43	<0.001	2, 48	5.70	0.006
Shape*alpha	6, 144	16.71	<0.001	6, 144	38.17	<0.001
Color*alpha	12, 288	58.95	<0.001	12, 288	10.86	<0.001
Alpha*shape*color	12, 288	6.16	<0.001	12, 288	5.75	<0.001

2.2.2 *Expt 1B: Assessing Visual Acceptability*

We then evaluated how "visually acceptable" the stimuli appeared, when varying their opacity across both emissive and AR displays. For emissive stimuli, opacity was induced across 7 levels of emissive alpha. For AR stimuli, opacity was induced across 7 levels of ΔL^*. Mixed effects logistic regressions accounting for repeated measures and binomial responses were conducted separately for emissive and AR displays. We present the extent to which stimulus opacity impacted visual acceptability, as well as the influence of the other independent variables (stimulus shape, stimulus color, background lightness), when statistically significant. Figure 6 summarizes the visual acceptability results from Experiment 1.

For stimuli viewed on the emissive display, there was a significant effect of opacity ($B = 15.08$, SE $= 1.79$, $z = 8.43$, $p < 0.001$), indicating that stimuli became more visually acceptable as they became more opaque. This effect did not significantly differ between light and dark backgrounds ($B = 1.04$, SE $= 1.21$, $z = 0.86$, $p = 0.39$). However, it is worth noting that the 50% acceptability threshold for light backgrounds was higher (emissive alpha $= 0.45$) than for dark backgrounds (emissive alpha $= 0.3$), suggesting that a greater range of transparency was more visually acceptable on dark backgrounds than on light backgrounds. Visual acceptability was not impacted by stimulus shape or stimulus color on emissive displays, $ps > 0.240$.

For stimuli viewed on the AR display, opacity had a significant effect ($B = 0.081$, SE $= 0.015$, $z = 5.23$, $p < 0.001$), indicating that stimuli became more visually acceptable as they became more opaque. Background lightness also had a significant effect ($B = 4.29$, SE $= 0.47$, $z = 9.21$, $p < 0.001$), indicating that visual acceptability was considerably lower when viewed on light backgrounds versus

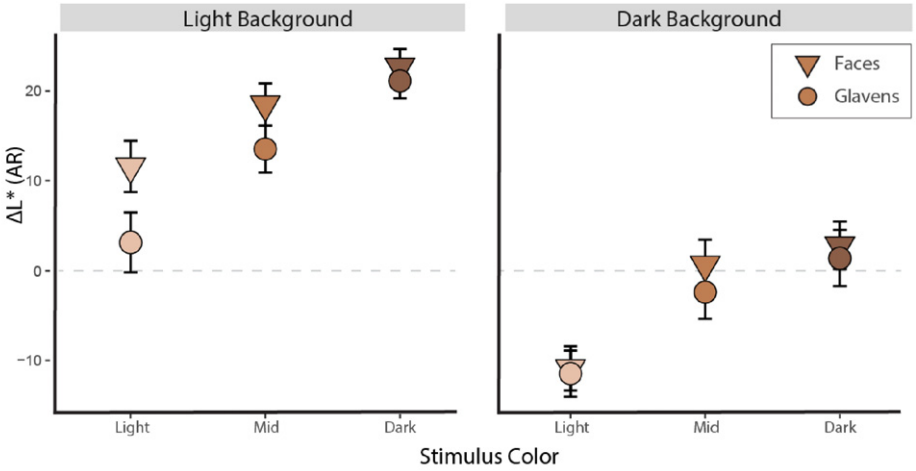

Figure 5. Summary of effects (mean and SE) of stimulus shape, stimulus color, and background lightness on L^* adjustments made for AR stimuli to match perceived transparency of emissive targets in Expt. 1. L^* needed to be increased more to match perceived transparency on the light versus the dark background, to match perceived transparency for the darker than the lighter stimuli, and to match perceived transparency for faces than glavens, which was particularly evident among lighter faces on the light background.

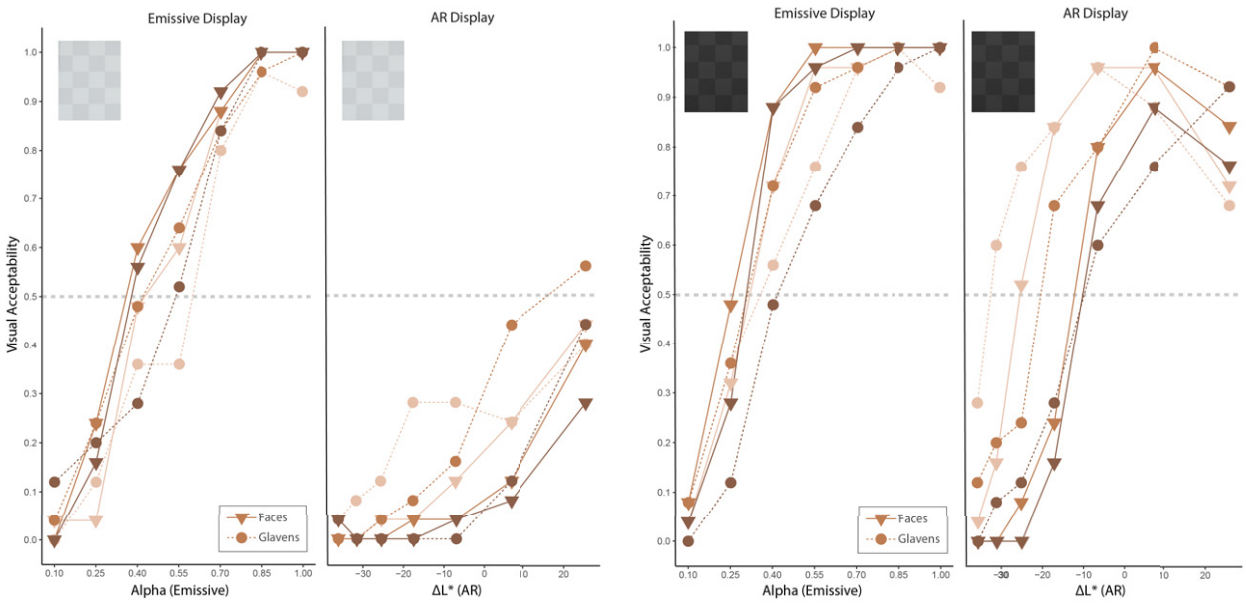

Figure 6. Visual Acceptability (y-axis) is shown for increasing levels of opacity on the emissive display (emissive alpha, left in each subplot) and AR display (ΔL^*, right in each subplot) over (left subplot) the light background and (right subplot) the dark background, as indicated by the plot embellishments. In both cases, higher levels of opacity had higher levels of visual acceptability, but opacity on AR displays had much lower acceptability than for emissive displays. On the dark background, AR acceptability decreases at the highest values due to the desaturation that occurs with high levels of lightness increase.

dark backgrounds. It is important to note that on average the stimuli did not exceed the 50% acceptability threshold at any ΔL^* (max $\Delta L^* = 26$) when viewed on light backgrounds, suggesting that adequately displaying AR objects in very bright environments is a particular challenge. Conversely, the 50% acceptability threshold for AR stimuli on dark backgrounds was relatively lower ($\Delta L^* = -17$), suggesting that some amount of transparency could still be considered acceptable in these conditions. There was a significant effect of stimulus shape ($B = 0.83$, SE $= 0.37$, $z = 2.23$, $p = 0.026$), indicating that faces were generally evaluated as less acceptable than glavens. In particular, increasing transparency was more detrimental to visual acceptability for faces than glavens ($B = -0.042$, SE $= 0.018$, $z = 2.29$, $p = 0.022$). The effect of stimulus color was significant, such that light stimuli were generally more visually acceptable than dark stimuli ($B = -1.06$, SE $= 0.51$, $z = 2.05$, $p = 0.04$), but not mid-tone stimuli ($B = -0.87$, SE $= 0.54$, $z = 1.63$, $p = 0.10$).

3. EXPERIMENT 2

This experiment explored the similarities and differences between transparency adjustments made to optimize

environmental fit and observer preference, in the context of a more realistic setting, under various lighting conditions.

The same OST-AR setup as in Experiment 1 was used, with the emissive display removed and a larger checkerboard background—using the same Michelson contrast as before, but with an average lightness midway between the previous light and dark checkerboards. The same stimulus set was used as in Experiment 1. Three different lighting conditions (cool, warm, magenta) were used, in order to create pairings of viewing booth lighting and simulated stimulus lighting which would be displayed in both matching and mismatching conditions. Mismatching conditions were hypothesized to potentially benefit from an increase in perceptual transparency (via a decrease in L^*) in order to allow more perceptual mixing of the background reflectance with the overlaid AR stimulus, resulting in a reduction of the perceived difference between stimulus-lighting and booth-lighting.

3.1 *Methods*

A printed checkerboard was used as wallpaper on the back of the viewing booth, to provide a contrast pattern consistent with the previous experiment without acting as an in-plane "canvas" for the stimuli. Several physical objects were placed into the booth: fake fruit of different colors (including glossy plastic grapes for their specular highlights which convey useful information about the lighting condition) and a miniature XRite ColorChecker. These objects were placed in the booth to aid observers in contextualizing the scene's lighting outside of the AR stimulus: the presence of familiar objects leverages memory color to improve color constancy; more broadly, increased complexity of cues to the illuminant present in visual stimulus has been shown to improve observers' ability to compensate for viewing condition changes [13, 17].

The three lighting conditions were calibrated to be equiluminant at ∼150 cd/m^2. They were the previously-used chromaticity match to D65 (cool), a chromaticity match to Illuminant A (warm), and a custom magenta with chromaticity [$x = 0.37$, $y = 0.24$] (see Figure 7). This magenta was chosen to probe the impact of lighting orthogonal to the Daylight/Planckian locus, with the pink direction being prioritized over green for its slightly less qualitatively-negative effect on skin tones. The room lighting was turned off for this experiment, to allow complete observer adaptation to the booth illuminant. Each stimulus was rendered as appearing under all three illuminants, transformed using the CAM16 chromatic adaptation transform, CAT16 [14].

3.1.1 *Procedure*

Observers were walked through a brief demonstration of the three lighting conditions across both matched and mismatched stimuli, before commencing the experiment. Trials were presented in blocks with the same booth lighting condition, with observers alternating between starting with either the cool or warm lighting conditions (randomly selected), and all observers ending with the magenta lighting condition (due to this condition being more unconventional). At the start of each lighting condition, observers adapted for 20 seconds to cool and warm to reach at least 90% adaptation [24], and for 60 seconds to magenta due to its unfamiliarity to observers.

Using the same L^* gamma adjustment described above in Section 2.1.1, observers were asked to adjust the images for each of two different tasks representing distinct design goals, "preference" and "environmental fit". For environmental fit: *"Adjust the stimulus until it looks like it fits within the environment and its lighting,"* and for preference: *"Adjust the stimulus until it looks as good as possible within the environment."* An abbreviated form of the instruction text was displayed within the booth, for observer's reference. In each lighting block, trials were grouped according to the task and repeated twice. Within each group, trial order was randomized between the different {stimulus, stimulus-illuminant} combinations. Across all illuminants and questions, participants completed a total of 216 trials.

Twenty-four observers participated in the experiment ($M_{age} = 28$, $SD_{age} = 7.73$). There were 12 women, 10 men, and 2 non-binary individuals. Participants reported their ethnicity as: 7 Asian, 1 Black or African American, 14 white, 2 Hispanic, 1 multiracial, 1 opted not to respond. No observers indicated having a color vision deficiency, and all had normal or corrected-to-normal vision.

3.2 *Results*

Repeated measures ANOVA was done to evaluate the extent to which participants adjusted stimulus lightness (which induced stimulus transparency in OST-AR) to optimize their preferred appearance, and perceived environmental fit. The independent variables included stimulus shape (2; faces versus glavens), stimulus color (3; light, mid-tone, dark), and whether or not the illuminant lighting condition and stimulus lighting condition matched (2; matched versus mismatched). The dependent variable was mid-scale ΔL^*, the variable used in Experiment 1. As discussed previously, increases in stimulus lightness tended to make the stimulus appear more opaque (less transparent). Figures 8 and 9 summarize the results of Experiment 2.

3.2.1 *Preference*

First, we evaluated the extent to which participants adjusted stimulus L^* to optimize their preferred appearance, i.e., how participants altered the stimulus to make it look as good as possible to them. There was a significant effect of stimulus shape on ΔL^*, $F(1, 23) = 5.19$, $p = 0.032$, indicating that participants generally increased L^* for glavens ($M = 10.94$, $SE = 1.48$) more than for faces ($M = 8.51$, $SE = 1.12$) to optimize their preferred appearance. There was a significant effect of stimulus color $F(2, 46) = 154.97$, $p < 0.001$. Participants increased L^* more for dark stimuli ($M = 14.42$, $SE = 1.27$) than for mid-tone stimuli ($M = 10.36$, $SE = 1.27$), $t(23) = 9.53$, $p < 0.001$, and increased L^* more for mid-tone stimuli than for light stimuli ($M = 4.39$, $SE = 1.20$), $t(23) = 11.06$, $p < 0.001$, to optimize their preferred appearance. However, these effects were qualified

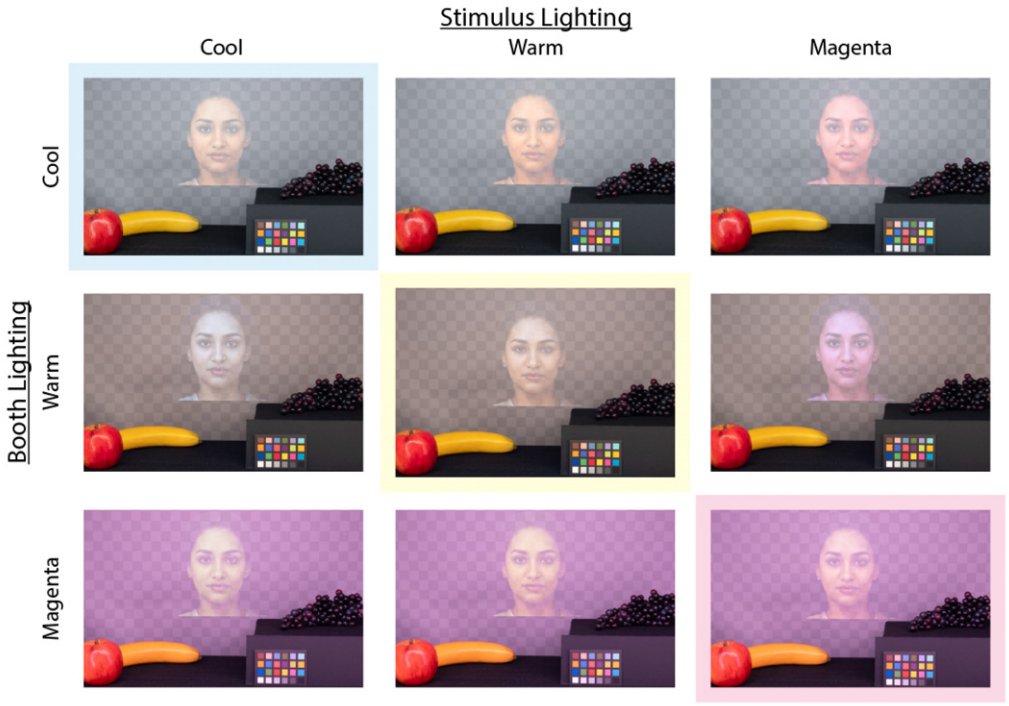

Figure 7. Photographs of combinations of booth and stimulus lighting used in Experiment 2. The three images along the diagonal are "matched" between the booth and stimuli, while the off-diagonal images are mismatched. Note that for illustration, the white balance settings of the images for each booth lighting setting were adjusted to approximate the visual effect of incomplete chromatic adaptation to the environment.

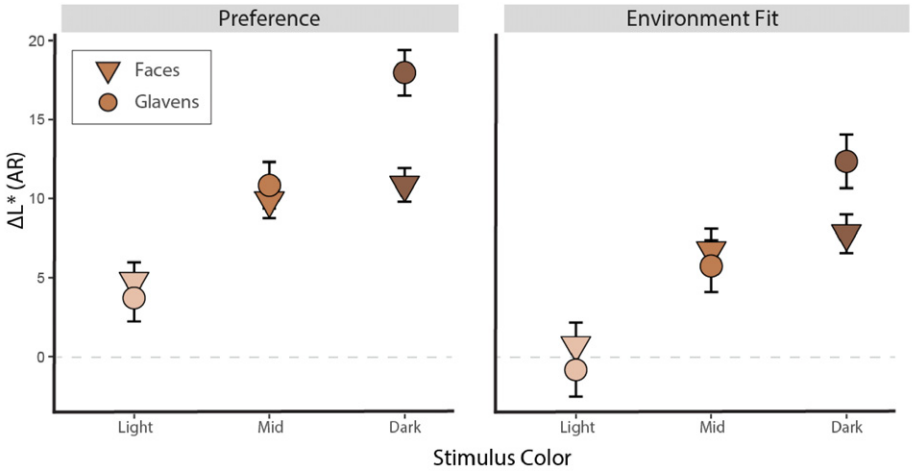

Figure 8. Summary of effect of stimulus shape and color on adjusted L^* determined in Experiment 2. Participants increased L^* more for dark stimuli, followed by mid-tone and light stimuli. The adjustments were consistent between faces and glavens, except for dark stimuli, where dark faces did not have comparable L^* adjustments to dark glavens.

by a significant stimulus shape*color interaction, $F(2, 46) = 64.63$, $p < 0.001$, indicating that this influence of stimulus shape on adjusted L^* varied as a function of stimulus color. Further exploring this interaction indicated that participants increased L^* more for glavens ($M = 17.95$, SE $= 1.53$) than for faces ($M = 10.89$, SE $= 1.19$) when stimuli were dark, $t(23) = 6.90$, $p < 0.001$. But, the ΔL^* differences between faces and glavens were not statistically significant for mid-tone stimuli, $t(23) = 1.01$, $p = 0.32$, or for light stimuli, $t(23) = 0.77$, $p = 0.45$, when optimizing their preferred appearance.

The effect of lighting match between illuminant and stimuli was not significant, $F(1, 23) = 0.94$, $p = 0.34$, indicating that the match (or mismatch) between booth-illuminant and stimulus-illuminant conditions did not generally impact ΔL^* when optimizing their preferred appearance. There were no additional significant interactions among these variables, $ps > 0.12$.

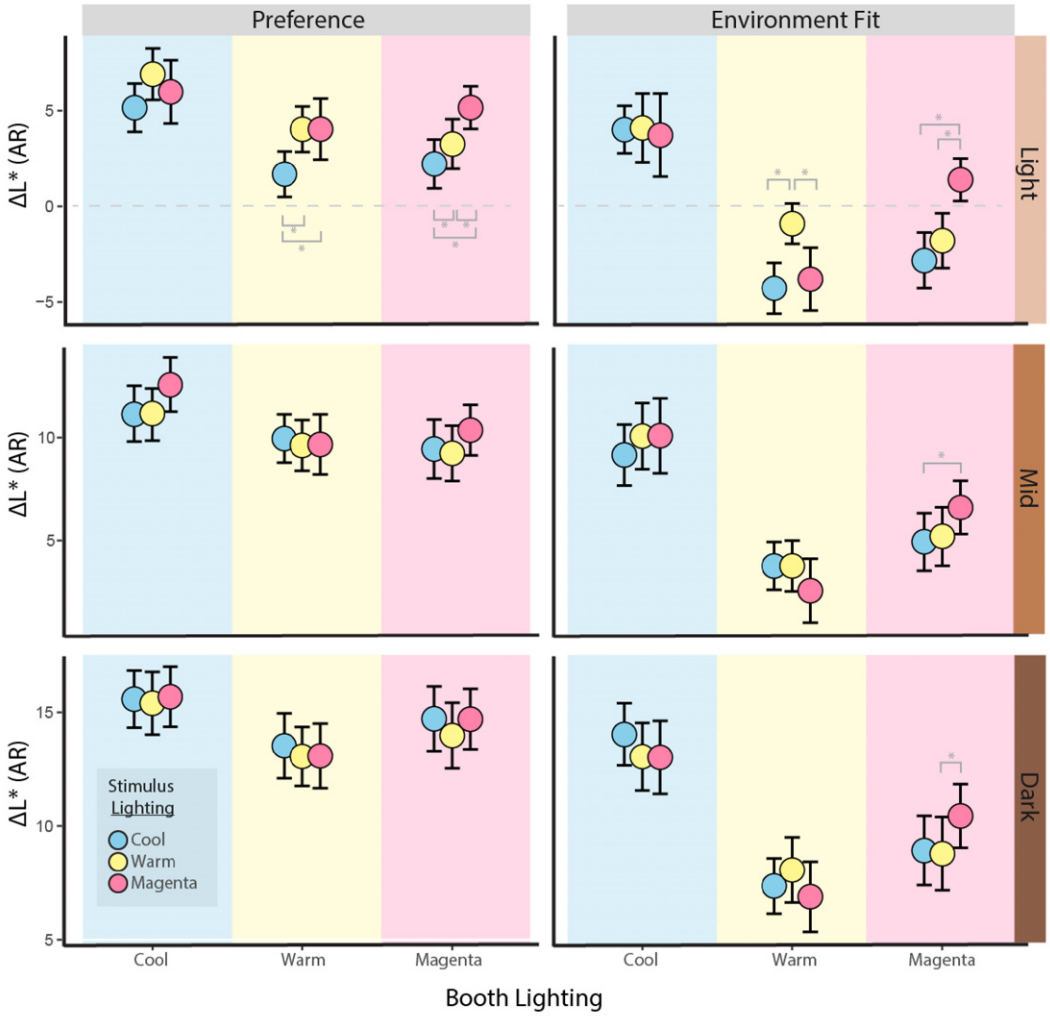

Figure 9. Summary of ΔL^* adjustments when illuminant and stimuli lighting conditions match versus mismatch (denoted here via match or mismatch in marker versus plot background color) in Experiment 2. Mismatched lighting conditions largely did not impact preferred appearance, except for some illuminants among light stimuli. The match between lighting conditions had more impact on environmental fit, particularly for light stimuli, and when the illuminant was magenta.

3.2.2 *Environmental Fit*

We evaluated the extent to which participants adjusted ΔL^* to optimize stimulus fit within the environment, i.e., how participants altered the stimuli to appear as if both the physical illuminant and AR stimulus were situated within the same environment. There was an insignificant effect of stimulus shape on adjusted ΔL^*, $F(1, 23) = 0.58$, $p = 0.45$, but a significant effect of stimulus color on adjusted ΔL^*, $F(2, 46) = 201.14$, $p < 0.001$, was seen. Participants increased L^* more for dark stimuli ($M = 10.26$, SE = 1.45) than for mid-tone stimuli ($M = 6.31$, SE = 1.45), $t(23) = 14.68$, $p < 0.001$, and increased L^* more for mid-tone stimuli than for light stimuli ($M = 0.34$, SE = 1.33), $t(23) = 11.11$, $p < 0.001$, to optimize their environmental fit. However, these effects were qualified by a significant stimulus shape*color interaction, $F(2, 46) = 62.79$, $p < 0.001$, indicating that this influence of stimulus shape on ΔL^* varied as a function of stimulus color. Further exploring this interaction indicated that participants increased L^* more for glavens ($M = 12.56$, SE = 1.84) than for faces ($M = 7.96$, SE = 1.21) when stimuli were dark, $t(23) = 4.13$, $p < 0.001$. But, the ΔL^* differences between faces and glavens were not statistically significant for mid-tone stimuli, $t(23) = 0.86$, $p = 0.397$, or for light stimuli, $t(23) = 1.12$, $p = 0.276$, when optimizing their environmental fit. These patterns are descriptively similar for both environmental fit and preference, but we note that the observed differences between these tasks were statistically significant, $F(2, 46) = 5.98$, $p = 0.005$.

The effect of lighting match between illuminant and stimuli was marginally significant, $F(1, 23) = 3.56$, $p = 0.072$, suggesting that participants somewhat tended to increase L^* more when lighting conditions matched ($M = 6.29$, SE = 1.26) versus mismatched ($M = 4.99$, SE = 1.58). However, this effect was also qualified by a significant lighting match*stimulus color interaction, $F(2, 46) = 8.70$, $p < 0.001$, indicating that the effect of lighting match varied as a function of stimulus color. Exploring this interaction

indicated that participants increased L^* more when lighting matched ($M = 1.50$, SE $= 1.09$) versus mismatched ($M = -0.81$, SE $= 1.65$) primarily for light stimuli, $t(23) = 2.67$, $p = 0.14$. For dark stimuli, the difference in ΔL^* between lighting match ($M = 10.85$, SE $= 1.38$) versus mismatch ($M = 9.67$, SE $= 1.59$) was only marginally significant, $t(23) = 1.85$, $p = 0.078$. For mid-tone stimuli, the difference was not significant, $t(23) = 0.59$, $p = 0.56$. No other significant interactions among these variables were found, $ps > 0.23$.

4. GENERAL DISCUSSION

Transparency for elements among conventional emissive displays can be adjusted via alpha-compositing, such that higher alpha values of an image appear more opaque. For OST-AR displays, transparency for virtual objects can be adjusted via the lightness of stimuli; the more light added to the AR system, the more opaque they appear. This was the approach used to manipulate perceived transparency for OST-AR objects. In Experiment 1, we had participants adjust AR-stimuli lightness in order to match the perceived transparency of the same stimuli viewed on an emissive display. We additionally evaluated how the range of transparency on both display types impacted stimulus appearance. In Experiment 2, we had participants adjust AR-stimuli lightness in order to make the stimuli appear most preferable and fitting within the environment, across three different lighting conditions (for the stimuli and for the viewing booth).

The results of this study provide a scale of perceived transparency in OST-AR that can be compared to transparency for more traditional emissive displays. This mapping may be useful in adjusting AR objects to predictably estimate their perceived transparency across different environments. As expected, we inferred that it became more difficult to match AR stimuli as the emissive stimuli became more opaque. This was particularly evident for darker stimuli, and when stimuli were viewed on light backgrounds. This was expected due to OST-AR's approach of additive light mixing; darker stimuli needed much more light added to be visible compared to lighter stimuli, which was exacerbated when the ambient light was already very high. In such conditions, it is likely that additional ambient light attenuation or rendering techniques are needed, such as dynamic occlusion or contrast boost as proposed in other recent studies [5, 8].

While assessing participants' perception of visual acceptability across the range of stimulus transparency, we found several notable differences between display types, stimuli, and backgrounds. For emissive displays, stimulus transparency predictably decreased acceptability, a pattern which did not vary substantially between different kinds of stimuli or backgrounds. This is likely because transparency on emissive displays was not induced by an additive light mixture, but rather an adjustment via the alpha-compositing of emissive stimuli and backgrounds.

Conversely, acceptability for AR stimuli was more nuanced. For AR stimuli on light backgrounds, acceptability was very poor across the full range of transparency, only approaching the "acceptable" threshold at the very high end of opacity used in the current work. Acceptability for AR stimuli was greatly improved when viewed on dark backgrounds, likely because substantially less AR light was needed to reproduce the stimuli in an appreciable way. Still, even in this condition, transparency became more detrimental to acceptability for darker stimuli than for lighter stimuli. Again, this was expected because darker stimuli needed additional light to adequately reproduce the objects than lighter stimuli. Additionally, transparency was more detrimental to acceptability for faces than for non-faces. This is possibly because faces represent a socially and cognitively meaningful stimulus; past studies have shown that people are particularly sensitive to facial skin color information [7]. The non-faces, on the other hand, represent a relatively "meaningless" (in a social-cognitive sense) object that could likely undergo more substantial changes in its appearance while still maintaining how it is evaluated. These findings agree with earlier studies [20] that demonstrate a current challenge for implementing social interaction in existing OST-AR technology, where adequately reproducing human faces, particularly those with darker skin tones, needs further research and improvements. An additional nuance for AR stimuli on dark backgrounds was that above a certain level of image lightness, acceptability decreased, indicating that there is an upper bound to lightness increase beyond which observers are no longer satisfied with stimulus appearance, regardless of shape.

We then assessed how AR transparency impacted perceptions of preference and environmental fit, among combinations of simulated lighting conditions for both AR stimuli and the viewing booth. Although observers adjusted lightness, the result in OST-AR is effectively an adjustment of perceived transparency—how much of the illuminant chromaticity "mixes" with the stimulus via the amount of background visible through the AR display. As a result, we anticipated observers using an adjustment in perceived transparency (via lightness) to mitigate perceptual mismatches in illuminant between stimulus and surrounding scene. As expected, participants tended to increase opacity more for darker stimuli than for lighter ones, both for preference and for environment fit. This was particularly problematic for darker faces, because increasing opacity via lightness likely resulted in faces appearing highly desaturated due to gamut limitations at high levels of lightness. This possibility might be supported by our findings that opacity for dark faces was not increased to the same level as dark glavens, for both preference and environmental fit. The large lightness increase needed for dark faces likely made them appear much worse than commensurate lightness level adjustments for the non-face objects. Again, this points toward the need for more research concerning the perception of diverse human faces in OST-AR.

Whether the lighting conditions of the AR stimuli and light booth matched (versus mismatched) did not have a substantial impact on preferred appearance, potentially

indicating that matching lighting conditions between the real environment and AR objects may not be a priority in OST-AR, if preferred appearance is the primary goal of the system. However, there may be an exception in the case of viewing lighter skin tones, where participants preferred more transparent stimuli when the lighting conditions mismatched in some cases. Matched (versus mismatched) lighting conditions had more of an impact on perceptions of environmental fit. Participants tended to make stimuli more transparent to fit the environment when the lighting conditions mismatched. This tendency was more evident when the light booth was set to our warm or magenta illuminants. This finding might indicate that when lighting conditions between AR objects and the real environment do not match, increasing AR transparency may be a useful strategy to facilitate a more coherent blending of the environment. It is worth noting that these differences tended to be most pronounced for light stimuli, which is contrary to most other findings described here, where we expected darker stimuli to pose the most challenges in OST-AR. On the other hand, it might have been the case that because lighter stimuli generally appeared more favorably, there was more flexibility to allow for more nuanced image alterations. Finally, perceptions for lighting matches (versus mismatches) were not as impacted when the light booth was set to the D65-approximating illuminant. We speculate that this might have been the case because people are generally more familiar with this lighting condition (e.g., most indoor artificial lighting approximates D65, and most standard image processing pipelines assume D65 as the default illuminant). However, the current experiment did not directly test this possibility.

Notably, we found substantial differences in transparency adjustments when evaluating preference versus environmental fit. Since these two tasks produced widely different outcomes, it is clear that the design goals of an AR system must be considered when determining the impact of stimulus properties on appearance. As such, tradeoffs in rendering and displaying AR objects may be necessary while considering the intended goals of the system. For instance, displaying AR objects with the goal of optimizing their apparent fit within the environment may consequently make them appear less preferable, and vice-versa. While this study only evaluated two distinctly probable design goals for AR systems (i.e., preference and environmental fit), other goals certainly exist. Yet, the observed findings point to the likelihood of other such trade-offs existing, prompting the need to consider how AR appearance impacts such goals independently.

This study had limitations that can be addressed in future studies in this area. First, while we were primarily interested in perceptions (and adjustments) of transparency, our method of altering AR transparency via lightness adjustments may have influenced color appearance, due to gamut constraints—particularly at very high levels of lightness. While we explicitly instructed participants to focus on perceived transparency (and ignore color appearance) it is still possible that color appearance contributed to some of the observations. Nevertheless, color gamut considerations are likely to be a practical limitation for scaling transparency in actual OST-AR systems, and so might be considered as an ecologically valid implementation. Our presentation of disembodied faces may not be fully ecologically valid, but is likely to be practically valid and relevant to our chosen use case of telecommunication, where users are often only visible above the shoulders. Finally, the current method uses a gamma adjustment on L^* to make stimuli appear more/less transparent. This mimics the additive effect of light, hence being justified here and in previous work [15]. However, given this method's likely impact on other aspects of appearance, such as color, this could certainly be improved, perhaps by introducing high dynamic range (HDR) capabilities to AR displays.

5. CONCLUSION

Augmented reality aims to combine elements of the real world and additional virtual objects together within the same viewing environment. OST-AR accomplishes this by adding light to the viewing environment using transparent optics. With this approach, the virtual objects will ultimately comprise some amount of transparency, particularly when those objects are darker, or are viewed in brighter real-world lighting conditions without additional modifications to the system. The current work aimed to characterize the perceived transparency of OST-AR objects by comparing them to a more familiar approach of alpha-compositing for emissive displays. Additionally, given that the lighting conditions of the real-world and virtual objects may often mismatch, we aimed to evaluate the extent to which this discrepancy might impede appearance when combined within an OST-AR environment. Further, because interpersonal communication is a particularly promising application for AR environments via teleconferencing, telehealth, and social contact, the current work focused largely on the appearance of human faces (versus non-faces) comprising different skin tones that correspond to variability across the human population. The results of this study provide a scale of perceived transparency in OST-AR as a function of stimulus and ambient conditions, and evaluated limitations of OST-AR appearance reproduction. These results may be useful toward predicting OST-AR transparency, and its potential impact on appearance. Finally, the current work highlights the need for additional research considering the interplay between AR and "real-world" environments, considering design goals in the development of AR systems, and improving the representation of human faces with respect to the diversity of skin tones across the human population.

ACKNOWLEDGMENT

This study is based on work supported by the National Science Foundation under Award No. 1942755.

REFERENCES

[1] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Facial color is an efficient mechanism to visually transmit emotion," Proc. Natl. Acad. Sci. **115**, 3581–3586 (2018).

[2] M. Doroodchi, P. Ramos, A. Erickson, H. Furuya, J. Benjamin, G. Bruder, and G. F. Welch, "Effects of optical see-through displays on self-avatar appearance in augmented reality," *EEE Int'l. Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (IEEE, Piscataway, NJ, 2022), pp. 352–356.

[3] T. Downs and M. J. Murdoch, "Color layer scissioning in see-through augmented reality," *Proc. IS&T CIC29: Twenty-Ninth Color and Imaging Conf.* (IS&T, Springfield, VA, 2001), pp. 60–65.

[4] J. P. Freiwald, N. Katzakis, and F. Steinicke, "Camera time warp: Compensating latency in video see-through head-mounted-displays for reduced cybersickness effects," *IEEE Int'l. Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (IEEE, Piscataway, NJ, 2018), pp. 49–50.

[5] C. Gao, Y. Lin, and H. Hua, "Optical see-through head-mounted display with occlusion capability," Proc. SPIE **8735**, 107–115 (2013).

[6] Generated Photos website, https://generated.photos.

[7] M. Hasantash, R. Lafer-Sousa, A. Afraz, and B. R. Conway, "Paradoxical impact of memory on color appearance of faces," Nat. Commun. **10**, 3010 (2019).

[8] J. D. Hincapié-Ramos, L. Ivanchuk, S. K. Sridharan, and P. Irani, "SmartColor: Real-time color correction and contrast for optical see-through head-mounted displays," *Proc. IEEE ISMAR* (IEEE, Piscataway, NJ, 2014), pp. 187–194.

[9] A. Ishihara, H. Aga, Y. Ishihara, H. Ichikawa, H. Kaji, K. Kawasaki, D. Kobayashi, T. Kobayashi, K. Nishida, T. Hamasaki, and H. Mori, "Integrating both parallax and latency compensation into video see-through head-mounted display," IEEE Trans. Vis. Comput. Graphics **29**, 2826–2836 (2023).

[10] R. E. Jack and P. G. Schyns, "The human face as a dynamic tool for social communication," Curr. Biol. **25**, R621–R634 (2015).

[11] B. C. Jones, L. M. DeBruine, J. K. Flake, M. T. Liuzza, J. Antfolk, N. C. Arinze, I. L. Ndukaihe, N. G. Bloxsom, S. C. Lewis, F. Foroni, and M. L. Willis, "To which world regions does the valence–dominance model of social perception apply?," Nat. Human Behav. **5**, 159–169 (2021).

[12] A. L. Jones, R. S. S. Kramer, and R. Ward, "Signals of personality and health: The contributions of facial shape, skin texture, and viewing angle," J. Exp. Psychol.: Hum. Percept. Perform. **38**, 1353–1361 (2012).

[13] J. M. Kraft and D. H. Brainard, "Mechanisms of color constancy under nearly natural viewing," Proc. Natl. Acad. Sci. **96**, 307–312 (1999).

[14] C. Li, Z. Li, Z. Wang, Y. Xu, M. R. Luo, G. Cui, M. Melgosa, M. H. Brill, and M. Pointer, "Comprehensive color solutions: CAM16, CAT16, and CAM16-UCS," Col. Res. Appl. **42**, 703–718 (2017).

[15] Z. Li and M. Murdoch, "Improving naturalness in transparent augmented reality with image gamma and black level," Proc. Color Imaging Conf. **30**, 147–152 (2022).

[16] A. C. Little, B. C. Jones, C. Waitt, B. P. Tiddeman, D. R. Feinberg, D. I. Perrett, C. L. Apicella, and F. W. Marlowe, "Symmetry is related to sexual dimorphism in faces: Data across culture and species," PLoS ONE **3**, e2106 (2008).

[17] S. Ma, R. Sun, Y. Liu, Y. Wang, and W. Song, "Effect of surrounding objects in the adapting scene on chromatic adaptation," Opt. Express **31**, 18587–18598 (2023).

[18] A. A. Michelson, *Studies in Optics* (University of Chicago Press, Chicago, 1927), p. 36.

[19] M. J. Murdoch, "Brightness matching in optical see-through augmented reality," J. Opt. Soc. Am. A **37**, 1927–1936 (2020).

[20] T. C. Peck, J. J. Good, A. Erickson, I. Bynum, and G. Bruder, "Effects of transparency on perceived humanness: Implications for rendering skin tones using optical see-through displays," IEEE Trans. Vis. Comput. Graphics **28**, 2179–2189 (2022).

[21] F. Phillips, E. J. L. Egan, and B. N. Perry, "Perceptual equivalence between vision and touch is complexity dependent," Acta Psychol. **132**, 259–266 (2009).

[22] F. Phillips, M. W. Casella, and E. J. L. Egan, Glaven objects (v1.4), 2016. Retrieved from github.com/skidvision/Glavens on July 20, 2023.

[23] M. G. Rhodes and J. S. Anastasi, "The own-age bias in face recognition: A meta-analytic and theoretical review," Psychol. Bull. **138**, 146–174 (2012).

[24] O. Rinner and K. R. Gegenfurtner, "Time course of chromatic adaptation for color appearance and discrimination," Vis. Res. **40**, 1813–1826 (2000).

[25] I. D. Stephen and D. I. Perrett, "Color and face perception," in *Handbook of Color Psychology*, edited by A. J. Elliot, M. D. Fairchild, and A. Franklin (Cambridge University Press, Cambridge, 2015), pp. 585–602.

[26] K. W. Tan and I. D. Stephen, "Colour detection thresholds in faces and colour patches," Perception **42**, 733–741 (2013).

[27] C. A. Thorstenson, A. D. Pazda, and A. J. Elliot, "Subjective perception of color differences is greater for faces than non-faces," Soc. Cogn. **35**, 299–312 (2017).

[28] C. A. Thorstenson and A. D. Pazda, "Facial coloration influences social approach-avoidance through social perception," Cogn. Emotion **35**, 970–985 (2021).

[29] L. Zhang and M. J. Murdoch, "Perceived transparency in optical see-through augmented reality," *2021 IEEE Int'l. Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (IEEE, Piscataway, NJ, 2021), pp. 115–120.

[30] G. Zonios, J. Bykowski, and N. Kollias, "Skin melanin, hemoglobin, and light scattering properties can be quantitatively assessed in vivo using diffuse reflectance spectroscopy," J. Invest. Dermatol. **117**, 1452–1457 (2001).