https://doi.org/10.2352/CIC.2024.32.1.23

Connecting Individual Color Matching Functions to Accuracy and Preference in Image Reproduction

Eddie Pei, Susan Farnand, Michael J. Murdoch; Munsell Color Science Laboratory, Rochester Institute of Technology; Rochester, New York 14623, USA

Abstract

Human perception of color varies between individuals, raising the question of how well the standard color matching functions (CMFs) represent individual observers in image reproduction. The goal of this research is to explore the relationship between CMFs and both fidelity and preference in the image reproduction pipeline. Three experiments were conducted: an experiment to estimate approximately individualized CMFs, an image fidelity experiment, and an image preference experiment. The results show that the CMFs influence the accuracy of image reproduction, however, preferences are affected by factors in addition to CMFs. The findings offer insights into the limitations and potential implications of relying on using standard CMFs in image reproduction technologies.

Introduction

Human vision exhibits significant variation among observers – not only in terms of color deficiency but also differences among people with color-normal vision. For instance, the eye's lens tends to yellow over time when exposed to ultraviolet radiation, significantly affecting sensitivity to shorter wavelengths. Other differences are caused by various factors including macular pigment density and shifts in cone peak sensitivity wavelength [1][10][24][25]. Alfvin and Fairchild designed and performed a visual experiment with color-normal observers to measure the magnitude of uncertainties associated with both intra- and inter-observer variability in cross-media color matching. They found that the average differences between the matches were 2.5 CIELAB units with peaks up to 19 CIELAB units. They also found the presence of age-related changes in color vision [2]. While effective cone sensitivities may be the physiological basis, color matching functions (CMF) are more directly measurable and essential for computing metamers.

Metamerism, a color phenomenon where spectrally different stimuli appear to be the same to a given observer under a given set of viewing conditions, can fail between observers who have differences in their vision systems. Observer metamerism became more evident with the introduction of narrow-band light sources in wide-gamut displays and LED lighting systems. The evolution of displays, such as OLED, laser displays, and others aiming for larger color gamuts through narrower primaries, accentuates the challenge of observer metamerism. This limitation has driven interest in the pursuit of personalized image reproduction. Canham et al. have explored a theoretical metameric match simulation colorimetric reproduction pipeline using different CMFs [8]. Kim et al. tapped into a deep-learning approach to personalized image enhancement [12]. In each of these, measuring the individual CMFs emerges as the most direct solution.

The measurement of CMFs can be traced back to early last century. Wright performed a CMF experiment in 1920s [26]. His data, along with J. Guild's color-matching experiment data and the luminous efficiency functions, helped to define the CIE 2° standard colorimetric observer in 1931 [20]. Subsequently, the 10° standard observer was developed based on data from 49 observers by Stiles and Burch [23] and 27 observers by Speranskaya [21] in the 1960s. Even today, these two standard sets of CMFs are well-used in many areas. Both standard sets of CMFs took the average of the observers' experimental data, and they did not consider the factors that cause the difference in color perception, meaning individual variation was not considered. Researchers have worked for decades to develop individualized CMFs. North and Fairchild designed a visual colorimeter to implement a new model for determining color-matching functions [15]. In 2006, the CIE published a technical report where the cone fundamentals for a population average can be calculated as a function of age ranging from 20 to 80 years old and a field of view ranging from 1° to 10° [5]. Asano, Fairchild, and Blonde summarised the many experiments that quantify observers' variability in 2016, and developed a statistical model of CMFs for individuals [4]. However, directly measuring the individual CMFs is troublesome, time-consuming, and requires specific instruments and a controlled environment. An alternative approach is to find categorical observers that represent clusters of the color normal population [3][18].

Building on these findings, the goal of our research was to test the potential for utilizing estimated individual CMFs into an image reproduction pipeline. Starting with a set of categorical CMFs based on previous research, we sought to rapidly estimate an individual's CMF category, then show them cross-media images reproduced using that set of categorical CMFs. In this research, the correlations among individualized CMF, fidelity, and preference in image reproduction were used to evaluate the congruence between CMF categories in each part of the study. Our expectations were that individualized color reproduction would have an effect on fidelity but might not have an effect on preference. The paper is organized as follows: Section 2 describes the techniques and stimuli used in the study; Section 3 details the three experiments—CMF determination, image fidelity, and preference—along with their respective findings; and the final sections discuss the results and their broader implications.

Experimental Preparation

Color Matching Functions

Differences in the visual system, even within the normal color vision population, can lead to variation in color perception. Factors such as macular pigment, lens density, and others

(a) The tri-pixel display (b) The viewing booth (c) The middle pixel's SPDs of primaries in tri-pixel display

Figure 1: Details of the lighting systems used in experiments

that affect color perception could be quantified by CMFs. Traditional methods for measuring CMFs are troublesome and time-consuming. Consequently, the CIE researched a more efficient approach for decades resulting in the CIE06 CMFs. This framework included the calculation of average cone fundamentals across a range of field sizes (from 1° to 10°) and ages (from 20 to 80). However, the CIE06 fundamental model is a theoretical construct based on averages, and the peak-wavelength shift factor is not taken into account in the model. Thus, the model is unable to predict any variation.

Building on the CIE06 model, previous work introduced a proposal for a set of categorical observers which could be applied in the quantification and prediction of observer metamerism. The idea was initiated by Sarkar et al. [18], where eight observer categories and their corresponding CMFs were introduced based on the classification of color-normal human observers. However, Sarkar's methodology has limitations: the relatively small sample size of 49 (and 47 were used) and a single field of view (10°). Therefore, the usage of these CMFs are constrained to specific conditions.

In response to these limitations, Asano and Fairchild [3] proposed a categorical observer model. The model first adopted a Monte Carlo statistical sampling to simulate 1000 CMFs and then applied a clustering method (k-medioids algorithm) to define a certain number of CMFs. This CMF model efficiently captures the diversity among the normal color vision population. It can be more flexible to the changing field size and the number of CMFs and is more robust to changes in spectral combinations for the use of different applications and aligning with the principles of CIE06. According to their evaluation, ten CMFs are sufficient to categorize observers. Thus, this study adopts Asano's ten 10° categorical observer CMFs. The characteristics of these category CMFs are detailed in Asano et al's work [3]. Among these ten-category observer CMFs, Category one is the same as CIE06 with age 38-year-old and a given field size. The second and third most significant categorical observers exhibit substantial deviations in lens pigment density and age. Categorical observer 10 is most representative of an older population. We added an 11th category, the CIE 1964 standard 10° CMFs, because it has been widely used. Overall, without *a priori* knowledge of any individual's CMFs, these categorical observers were chosen to be potentially representative of the general population of color-normal observers.

Lighting System

Two lighting systems were utilized in separate experiments, as shown in Figure 1: A tri-pixel display and a viewing booth. The tri-pixel display (Figure 1a) consisted of three separate "pixel" subsystems. Each system had an LED lighting system and a diffuser. The LED lighting system was an ETC Source Four LED seven primary spotlight designed for theater use. Each primary was addressed in a channel via an 8-bit (0-255) digital signal. Similarly, the viewing booth, shown in Figure 1b, contained two independent subsystems. The advantage of the LED system was that, given a lighting condition, it could match the reference reasonably. These lighting systems were evaluated by Yuan et al. [28] regarding precision and accuracy. In the current work, the spectral power distributions (SPDs) of the primaries within both lighting systems were measured by using a PR-655 spectroradiometer. The measurements of the viewing booth also required a Polytetrafluoroethylene (PTFE) white reference sample. The measured middle pixel's SPDs of primaries in the tri-pixel display are shown in Figure 1c; the SPDs of the light booth consist of the same LED primaries and are, consequently, not shown separately.

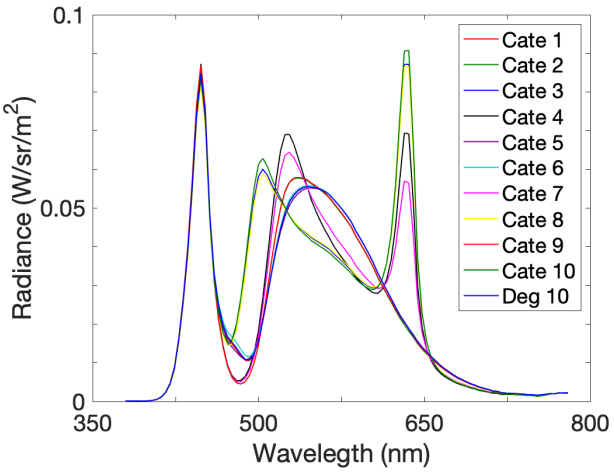


Figure 2: The SPDs of the stimuli in this experiment.

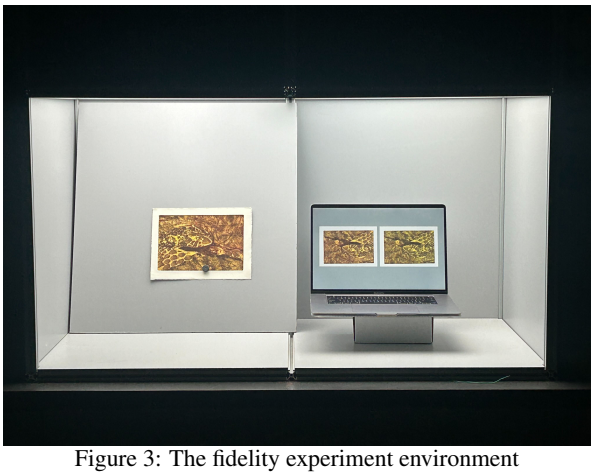


Figure 3: The fidelity experiment environment

Stimuli

Observer metamerism is one of the core concepts in this study, which incorporates 11 different CMFs. Two types of stimuli were created: the lighting metamer pair stimuli and the image metamer pair stimuli, detailed below.

Light Stimuli

The lighting metamer pair stimuli were presented on the tripixel display, whose LED light system is capable of approximating an aim spectral power distribution SPD_{aim} by applying different weights (β_{weight}) to each of the seven LED primaries. This adjustment was guided by the following equation:

$$SPD_{aim}(\lambda) = \sum_{i=1}^{7} \beta_i \times SPD_i(\lambda) + S(\lambda)_{flare} \tag{1}$$

The MATLAB 'fmincon' function was applied to determine the weighting vector $[\beta_1 \dots \beta_7]$ with an offset term in case of nonzero flare $S(\lambda)_{flare}$. First, an SPD of D65 was selected, and the seven primaries of the middle pixel were used to create the reference stimulus, denoted as SPD_{D65}. Subsequently, fewer LED primaries on the left and right pixels were used to generate metameric pairs for each of the 11 observers' CMFs.

To create metamers, either three (35 combinations) or four (35 combinations) primaries on the left pixel were used to generate its SPD. The calculation procedure was similar to Murdoch's work [14]. The SPD generation was constrained to match the tristimulus (XYZ) values, which were calculated by each category observer's CMF, while simultaneously maximizing spectral difference relative to the target SPD_{D65}. For the right pixel, the computation was repeated, aiming to replicate the SPD of the left pixel. Thus, 22 pairs of metamers were created. The final results are shown in Figure 2.

A similar method was adopted for the viewing light booth to match a D65 lighting illumination for the two sides of the lighting system. The SPD of this light is denoted as SPD_2.

Image Stimuli

Starting with original paintings (hard-copy) shown in Figure 4, a set of soft-copy images was created to be metameric for each of the categorical CMFs. Cross-media image comparison is a relevant imaging task, and it relates to a recent cross-media color-matching experiment that has been validated as a swift and reliable method for identifying superior performing individual CMFs [19]. The selected paintings each contain, at most, a few hues because too many different hues within a single image can make it challenging for observers to make consistent judgments. Images 1 and 4 are mostly orange and brown. Image 2 includes green, blue, and purple, while Image 3 contains blue and yellowish hues. This section outlined the methods of converting hard-copy images into soft-copy images with different category observers' CMFs. These transformations were performed using the hard-copy image's estimated reflectance properties, as determined by a camera-based model, reproduced using individual CMFs.

Spectral estimation using a camera-based model is traditionally employed in the field of cultural heritage [7] [13], as well as characterization, verification, and documentation of the stimuli in augmented reality (AR) displays [9]. In order to generate the image material pair stimuli, a comparable approach is adopted. The spectral estimation imaging process involves the use of a commercially available RGB camera along with two color filters, namely, one yellow and one cyan, to capture a 6-channel image array. Another component was a known spectral training set, such as color targets. This process is detailed in Kuzio et al.'s work [13]. The underlying model for this method is shown as follows:

$$\begin{pmatrix} R_1 \\ G_{1,1} \\ B_1 \\ R_2 \\ G_{1,2} \\ B_2 \end{pmatrix} \xrightarrow{\widehat{f}} \begin{pmatrix} \lambda_{380} \\ \lambda_{390} \\ \vdots \\ \lambda_{730} \end{pmatrix} \tag{2}$$

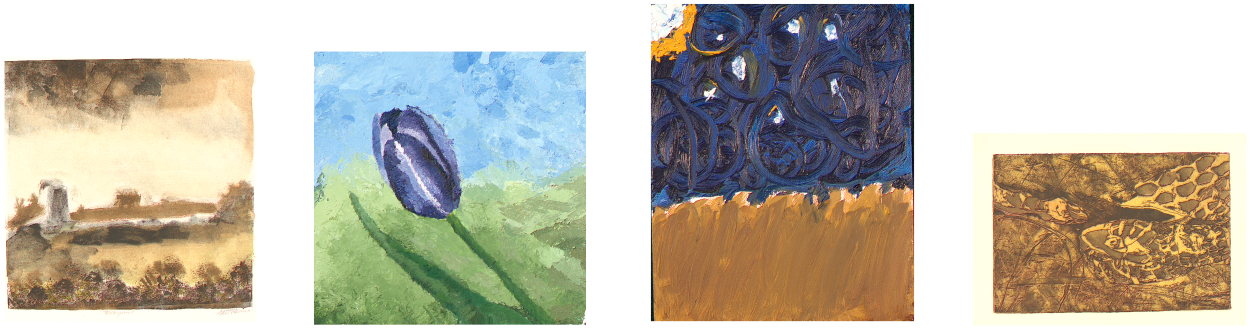
The goal was to determine the $\widehat{f}$. The loss function was the mean squared error of the estimated and the known spectral reflectance. Then, when inputting a new image, the $\widehat{f}$ could be used to estimate the reflectance properties.

For this purpose, a Canon EOS 5D Mark II camera was used to capture two sets of three output values per pixel under one yellow and one cyan-filtered lighting of four color targets, namely Next Generation Target (NGT), Artist Paint Target (APT), Color Checker Semi-Gloss (CCSG), and Color Checker classic (CC classic). The reflectances of the color targets and the output image were used to train the model($\widehat{f}$), which was then used to perform spectral estimation on the images. The methodology including the two filters and color targets were the same as that used by Kuzio [13].

The spectral reflectance (R_{image}) estimations of four image scenes were conducted by the ($\widehat{f}$). Following this process, the tristimulus values were calculated using the 11 CMFs and the SPD of the viewing light booth (SPD_2), with a wavelength step of $\Delta\lambda$. For each CMF (CMF_i), the corresponding tristimulus value is represented as XYZ_{CMFi}. As shown in equation 3

$$XYZ_{CMFi} = CMF_i \times SPD_2 \times R_{image} \times \Delta\lambda \tag{3}$$

A MacBook Pro served as the display for the experiment. Initially, 11 individualized display characterization models were developed based on the display model [6]. The transformation from tristimulus values to display RGB values was guided by the conceptual equation (including both linear and nonlinear components of the



(a) Image 1 (b) Image 2 (c) Image 3 (d) Image 4

Figure 4: The images used in the experiment. They were produced by Category One observer's CMF and SPD_2. These images are from Studio for Scientific Imaging & Archiving of Cultural Heritage lab.

display model):

$$XYZ_{CMFi} \xrightarrow{f_i} RGB_{Mac} \tag{4}$$

where the f_i is the display model that were created with the category CMF_i instead of standard observer CMF. Given the XYZ_{CMFi}, the soft copy was created by equation 4. The soft-copy images presented in Figure 4 were generated using Category One's CMFs.

Experimental Methodology and Results

Participants

Eight observers with normal color vision participated in the experiments. Each observer completed the Ishihara Color Vision Test. Each provided informed consent after receiving the procedural details. RIT's Human Subjects Research Office has approved this experiment. Among the eight observers, six had prior experience with color experiments, while the other two were naive observers. Four were in their 20s, three were in their 30s, and one was in their 40s.

Part I: Observer Categorization

CMF Experiment Setup

The objective of the CMF experiment was to determine the approximate CMFs of each observer efficiently. This approach is fundamentally based on observer metamerism. Yuan did a simulation study of observer metamerism by different CMFs [27]. Chen used mixed paint samples to match a display to determine observers' CMF [19]. In this work, we used a relatively simple sample - LED light metamer pairs. The light metamer pair stimuli listed in Figure 2 were used. The experimental procedure was as follows: the study was conducted in a room with average lighting conditions, where observers sat two meters from the tri-pixel display. Each pixel is 32 by 32 cm, and the Field of View (FOV) of each pixel is approximately 9°. A GUI created in Matlab was used to control the experiment. Before the test, a 15-minute warm-up period was required for the tri-pixel display, followed by a two-minute adaptation time for observers. The middle pixel stayed constant throughout the experiment, serving as the reference. The side pixels exhibited distinct metameric stimuli in a random order. A two-alternative forced-choice (2AFC) methodology was used, in which observers were tasked with choosing which of the left or right pixels more accurately matched the central reference. 55 pairs $C(11,2)$ were shown, and each observer completed five replications for a total of 275 trials (55×5). The expected duration of the experiment was approximately 20 to 30 minutes. The experiment setup is shown in Figure 1a.

Categorical CMFs Results

This experiment aimed to provide an estimate of an individual's color vision based on which metamers appear to match best. The results of the CMF experiment were obtained by the 2AFC method. A set of paired-comparison data, including 2200 observations (275×8), was collected. The data was converted to an interval scale by following Thurstone's Law of Comparative Judgment (Case V) [11]. The probit model was used to fit this data with an assumption of a normal distribution. The standard scores (z-scores) were derived from the probit model by transforming the probability values back into scores that correspond to the normal distribution under the probit model assumption. The same procedure was applied to process data in two subsequent experiments. The results were an interval scale as defined by Stevens' scale classification [22].

The interval z-score values essentially indicate how similarly an observer matches colors compared to each categorical CMF, where a higher z-score indicates a better correspondence to a given categorical CMF. Figure 5 column one presents eight observers' CMFs related to the categorical observers. Most observers from this experiment could be best represented by two to four CMFs, which may indicate that some of these category CMFs are correlated. For instance, Observer 1 aligned closely with Categories one and eleven, while Categories two and ten were less representative of this observer. In the case of Observer 6, categories seven and nine were the best fits, whereas categories three and ten were the least compatible. It is evident that Observers 1 and 6 in this experiment were represented by different CMFs, highlighting the individual variances. The same held true for other observers; unique results were observed for each. However, there are some commonalities. Notably, Category ten tended to be a less accurate representation for a majority of observers. Category ten was designed for an older demographic, whereas the age range of observers in this experiment was from 23 to 45. Category Observers three and seven have a large age gap, with opposing directions of lens and macular density relative to the average distri-

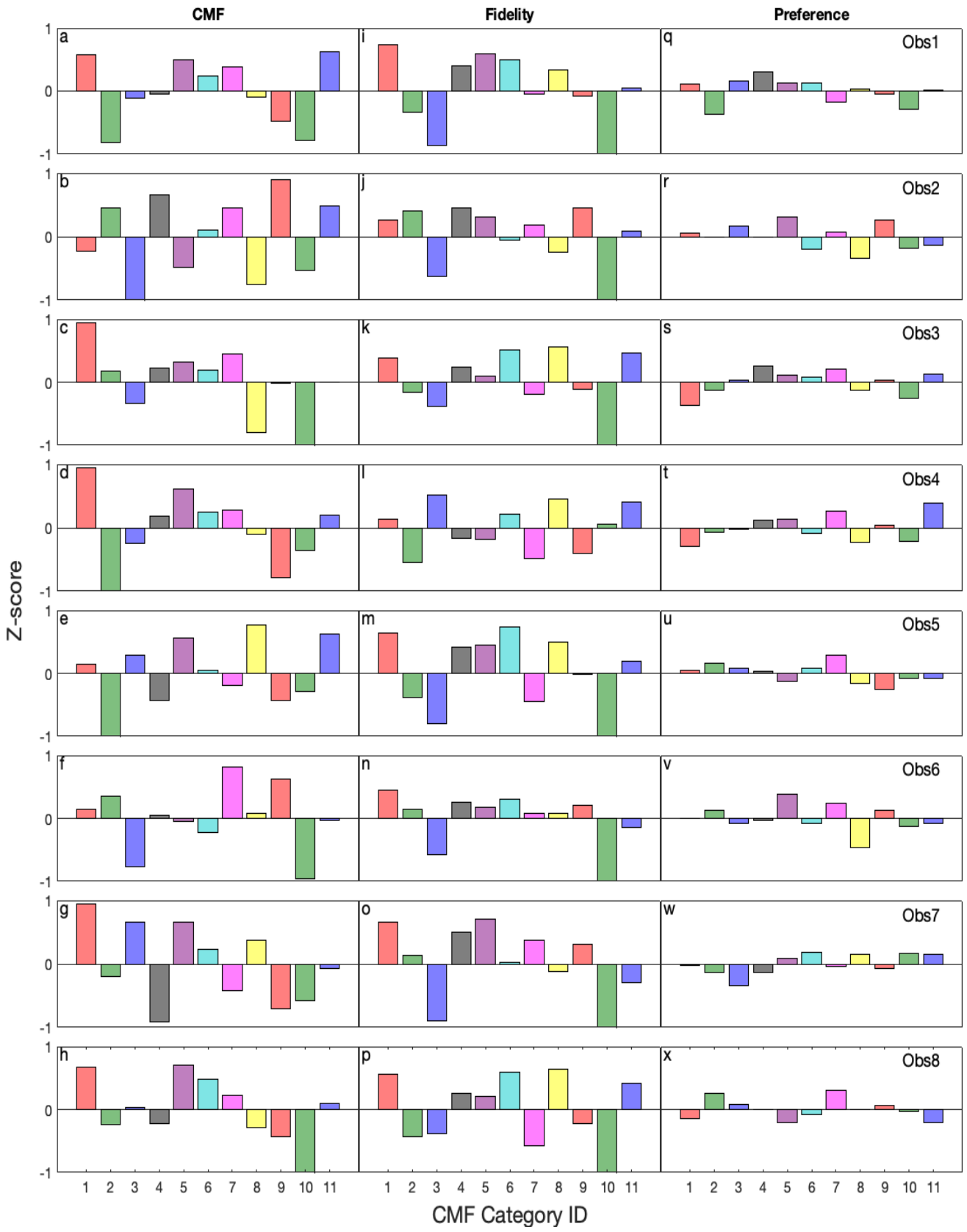


Figure 5: Three experiments results. For each subfigure, the x-axis represents the category observer ID from 1 to 11, and the y-axis shows the z-scores with the y-limits set to [-1, 1]. Note that the lowest z-score is -1.4, which is in Figure n. A higher score indicated that the observer match/preferred a given category of CMF's image reproduction.

Table 1: The background match experiment results. The RGB values are the encoded display RGBs, and cell background colors are rendered in sRGB to give an approximate visual representation of the matches. The $L^*a^*b^*$ and ΔE_{00} are computed from the CIE1964 10° CMF's tristimulus values. ΔL is the lightness difference. $\Delta E'_{00}$ is a modified ΔE_{00} that took out the lightness effect.

Obs ID	RGB	$L^*a^*b^*$	ΔE_{00}	ΔL	$\Delta E'_{00}$
Reference	-	(74.68 -0.02 -0.56)	-	-	-
Obs 1	(180 187 173)	(75 -5 6)	8.3	0.3	8.3
Obs 2	(184 179 168)	(73 0 6)	6.0	-1.7	5.9
Obs 3	(198 200 185)	(80 -3 7)	8.4	5.3	7.5
Obs 4	(140 149 145)	(61 -4 1)	12.2	-13.7	5.5
Obs 5	(156 159 145)	(65 -4 7)	11.0	-9.7	8.1
Obs 6	(179 180 170)	(73 -2 5)	5.8	-1.7	5.7
Obs 7	(153 150 143)	(62 0 4)	10.9	-12.7	4.3
Obs 8	(182 193 181)	(77 -6 5)	8.8	2.3	8.7

bution. Consequently,it would be unlikely for individual scores to be similar for the CMFs for these Observers, which corresponds with the contrasting z-score signs in our participant data. Without the knowledge of the observers, these observations from the CMF experiment indicated the variation of characteristics of these selected CMFs.

Part II: Image Preference

Background Color Matching Experiment

Prior to initiating the experiment, participants were given a 2-minute adaptation period to the illuminated light booth in an averagely lit room. Then, they were instructed to perform a color-matching task between the display background and the grey backdrop of the light booth. The booth area they matched was the location where the hard copy image was presented in the later experiment. This task was carried out through a MATLAB GUI. Participants adjusted the color in the CIELAB color space, which was then converted to RGB values and displayed on the screen. The experiment took about 5 minutes. The radiance of the chosen RGB of the display was measured using a CR250 radiometer, and the results were incorporated into subsequent phases of the study. The RGB values obtained from this color-matching task were used as the background settings for the following experiments.

Background Color Matching Results

The color matching experiment of the background served two primary objectives. First, it illustrated the variations in color perception among individuals. Second, the results were used as a background for the next two experiments.

The results are represented in terms of display RGB values and CIELAB values calculated using the 10° standard observer. The ΔE_{00} denotes the difference between an observer's result and the reference, as detailed in Table 1. Notably, the CIELAB values suggested that certain observers, like Observers 3 and 8, perceived the color to be brighter compared to others. In contrast, Observers 4 and 7 found it to be darker. The difference in lightness reached up to 18 units. This might be due to the non-uniformity of the lighting along the top to the bottom of the backboard, even though the observer was instructed to look at the same area where the hard-copy images were presented in the fidelity experiment. Due to eye movement and some uncontrolled variables, the perceived lightness had a larger variation. Chromatic difference in metameric pairs had been studied Park et al [16]. To study the chromatic attributes, the modified ΔE_{00} were calculated, which only considered the chromatic attributes, denoted as $\Delta E'_{00}$.$\Delta E'_{00}$ is calculated as follow:

$$\Delta E'_{00} = \sqrt{\left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T\left(\frac{\Delta C'}{k_C S_C}\right)\left(\frac{\Delta H'}{k_H S_H}\right)} \quad (5)$$

The notations have the same meaning as in ΔE_{00}. The results of $\Delta E'_{00}$ are presented in Table 1. The observers were classified as high-$\Delta E'_{00}$ differences, which include Observer 1, 3, 5, and 8, and low-$\Delta E'_{00}$ differences, which include Observer 2, 4, 6, and 7. Among the high-$\Delta E'_{00}$ group, three out of four results indicated that Category One best represented their color vision, and they all ranked Category Five within the top three. However, no clear trend was observed within low-$\Delta E'_{00}$ group

Preference Experiment Procedure

The experiment was conducted in an averagely lit environment where the observer was 1.5 meters away from the display, and the display was placed in the center of the viewing light booth(Figure 1b). The display and lighting were warmed-up for ten minutes prior to the experiment, and observers were allowed a 2-minute adaptation period. The stimuli were created using the first three image scenes in Figure 4. The fourth image scene has a similar color tint to the first one, hence, it was not used to reduce the experiment time. So, there were three different image scenes and 11 CMFs that created 33 pairs of image metamer stimuli. The soft-copy stimuli were designed to be around 13 by 13 ($\pm$1) cm, the FOV is 5°. The experiment was controlled via a GUI. A 2AFC method was adopted. All possible comparison pairs were considered. Therefore, for a given image, there are 11 metamer pairs, and there are 55 comparison pairs $C(11,2)$. During each trial, the two soft-copy images were displayed simultaneously, without the hard-copy painting present, and the observer was asked to select the one that most closely matched their personal preference. There was a 1-second pause showing the grey background between trials and no time limit to make a choice for observers. The experiment comprised two replicates, totaling 330 trials ($55 \times 3 \times 2$), and had a duration of approximately 20 minutes.

Preference Experiment Results

Image preference could be affected by many factors, for instance, memory color, cultural influences, and others [17]. In this

Table 2: Rank order of z-scores for three experiments: CMF, Fidelity, and Preference. Ranks are 1-11, where a smaller number means a higher z-score (better alignment with an observer's results), and a larger number means a lower z-score. For a given Category observer, the bold text with underlined text values highlights the best performers for observers when they agree across experiments. Red text color highlights the worst performers across experiments. Blue text color shows where results across experiments strongly disagree. The Correlation Coefficients of three experiments' z-scores are listed in the last three columns, along with their corresponding p-values. The null hypothesis is that there is no relationship between the test data sets. With α for p value set at 0.05, the significant correlations are marked ** under the p value.

Experiment Details		Categorical Observer											Correlation Coefficients		
Obs ID	Experiment	1	2	3	4	5	6	7	8	9	10	11	$corr_{C-F}$	$corr_{C-P}$	$corr_{P-F}$
Obs1	CMF	**2**	11	8	6	**3**	5	4	7	9	10	1	0.68	0.58	0.575
	Fidelity	**1**	9	10	4	**2**	3	7	5	8	11	6	$p = 0.02$	p =0.06	p =0.06
	Preference	5	11	2	1	3	4	9	6	8	10	7	**		
Obs2	CMF	7	4	11	**2**	8	6	5	10	**1**	9	3	0.69	0.15	0.41
	Fidelity	5	3	10	**1**	4	8	6	9	**2**	11	7	p =0.019	$p = 0.65$	p =0.21
	Preference	5	6	3	7	1	10	4	11	**2**	9	8	**		
Obs3	CMF	**1**	6	9	4	3	5	2	10	8	11	7	0.53	0.19	0.25
	Fidelity	**4**	8	10	5	6	2	9	1	7	11	3	$p = 0.09$	p =0.58	p =0.46
	Preference	11	8	6	1	4	5	2	9	7	10	3			
Obs4	CMF	1	11	8	6	2	4	3	7	10	9	5	0.29	0.076	-0.22
	Fidelity	5	11	1	7	8	4	10	2	9	6	3	p =0.38	p =0.82	p =0.51
	Preference	11	7	6	4	3	8	2	10	5	9	1			
Obs5	CMF	5	11	4	9	3	6	7	**1**	10	8	2	0.37	-0.38	-0.21
	Fidelity	2	8	10	5	4	1	9	**3**	7	11	6	p =0.26	p =0.25	$p = 0.54$
	Preference	5	2	3	6	9	4	1	10	11	8	7			
Obs6	CMF	**4**	3	10	6	8	9	1	5	**2**	11	7	0.74	0.40	0.26
	Fidelity	**1**	6	10	3	5	2	7	8	**4**	11	9	p =0.009	p =0.22	p =0.44
	Preference	5	3	8	6	1	9	2	11	4	10	7	**		
Obs7	CMF	**1**	7	2	11	**3**	5	8	4	10	9	6	0.1	0.009	-0.6
	Fidelity	**2**	6	10	3	**1**	7	4	8	5	11	9	0.76	0.98	0.86
	Preference	6	9	11	10	5	1	7	4	8	2	3			
Obs8	CMF	**2**	8	6	7	1	**3**	4	9	10	11	5	0.6	-0.32	-0.58
	Fidelity	**3**	9	8	5	6	**2**	10	1	7	11	4	p =0.05	p =0.33	p =0.06
	Preference	9	2	3	5	10	8	1	6	4	7	11	**		

experiment, the main focus was the color influence on the preference. There were 2640 observations in this experiment. The evaluation compared the z-scores, which were conducted following the same procedure described in the CMF experiment, as shown in Figure 5 column three. Observers 2 and 6 had similar preferred images, which aligned with the Category Five image reproduction. Category Seven's image reproduction was preferred by observers 4, 5, and 8. However, no image reproduction was preferred by all the observers. The preference varied among all observers. As shown in Figure 5, the z-scores are relatively lower, which indicates weaker or inconsistent preference of each category CMF's image reproduction. No clear relationship was identified between preference and CMF experimental results.

Part III: Image Fidelity

Fidelity Experiment Procedure

The experiment, conducted in an environment lit to around 260 lux measured at the location of observers' eye level, consisted of both the original hard-copy images and the soft-copy reproductions shown simultaneously within the light booth. The setup is shown in Figure 3. Observers were 1.8 meters away from the center of the screen and hard-copy painting. The FOV of the soft-copy stimuli is around $4.2 \pm 0.5°$. The widths of the hard-copy images are 19 to 32 cm, and their FOVs are $8 \pm 3°$. Both the light booth and display require a 10-minute warm-up period, followed by a 2-minute adaptation period for the observer. A 2AFC method was used with a Matlab GUI. Two images were displayed at a time, and observers were asked to select the image more closely resembling the hard-copy version. The experiment included four image scenes. Each image scene was used to create 11 soft-copy stimuli, resulting in 55 pair comparison trials for each image scene. The experiment was repeated twice, resulting in 440 trials for each observer. The experiment lasted approximately 40 minutes.

Fidelity Experiment Results

The evaluation of the fidelity experiment outcomes focused on identifying the best and worst-performing Category CMFs and comparing these with the CMF experiment results. These results are shown in Figure 5.

The results showed that Category Ten CMF's reproductions were a bad representation for almost all of the observers. This is the same as in the CMF experiment. The observers in this experiment were younger, therefore, this result was expected. Aside from this similarity, the rest of the results are individualized. For Observer 5, categories two, seven, and two CMF's produced in-

Table 3: The correlation and the p-values of the z-scores of the four individual images and the average across images (marked *All*), for the Fidelity experiment. For each observer, the correlations are presented in the top-right of the matrix and the p-values on the bottom-left.

Details		Correlation Coefficients and p-values				
Obs ID	Image	Image_1	Image_2	Image_3	Image_4	Image_{All}
Obs1	Image_1		0.095	0.85	0.86	0.90
	Image_2	p =0.78		0.12	-0.03	0.41
	Image_3	$p < 0.001$	p =0.72		0.87	0.91
	Image_4	$p < 0.001$	p =0.92	$p < 0.001$		0.87
	Image_{All}	$p < 0.001$	p =0.22	$p < 0.001$	$p < 0.001$	
Obs2	Image_1		-0.035	-0.0048	0.73	0.43
	Image_2	p =0.92		0.93	0.51	0.86
	Image_3	p =0.99	$p < 0.001$		0.596	0.89
	Image_4	p =0.01	p =0.11	$p = 0.05$		0.86
	Image_{All}	p =0.19	$p < 0.001$	$p < 0.001$	$p < 0.001$	
Obs3	Image_1		-0.34	0.85	0.73	0.75
	Image_2	p =0.3		-0.05	0.11	0.31
	Image_3	$p < 0.001$	p =0.88		0.92	0.91
	Image_4	p =0.01	p =0.75	$p < 0.001$		0.93
	Image_{All}	p =0.01	p =0.36	$p < 0.001$	$p < 0.001$	
Obs4	Image_1		-0.95	0.86	0.97	0.93
	Image_2	$p < 0.001$		-0.77	-0.96	-0.82
	Image_3	$p < 0.001$	p =0.01		0.88	0.98
	Image_4	$p < 0.001$	$p < 0.001$	$p < 0.001$		0.93
	Image_{All}	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	
Obs5	Image_1		0.71	0.77	0.87	0.95
	Image_2	p =0.01		0.78	0.44	0.81
	Image_3	p =0.005	p =0.005		0.79	0.93
	Image_4	$p < 0.001$	p =0.18	p =0.004		0.88
	Image_{All}	$p < 0.001$	$p = 0.003$	$p < 0.001$	$p < 0.001$	
Obs6	Image_1		0.09	0.06	0.49	0.75
	Image_2	p =0.80		0.96	-0.66	0.69
	Image_3	$p = 0.85$	$p < 0.001$		-0.72	0.65
	Image_4	p =0.13	p =0.03	$p = 0.01$		0.01
	Image_{All}	p =0.01	p =0.02	p =0.03	p =0.97	
Obs7	Image_1		0.60	0.46	0.90	0.89
	Image_2	p =0.05		0.82	0.57	0.89
	Image_3	p =0.16	$p = 0.001$		0.26	0.76
	Image_4	$p < 0.001$	$p = 0.07$	$p = 0.44$		0.81
	Image_{All}	$p < 0.001$	$p < 0.001$	$p = 0.01$	$p < 0.001$	
Obs8	Image_1		0.14	0.81	0.74	0.93
	Image_2	$p = 0.68$		0.26	-0.26	0.22
	Image_3	p =0.002	$p = 0.45$		0.91	0.86
	Image_4	p =0.01	$p = 0.44$	$p < 0.001$		0.84
	Image_{All}	$p < 0.001$	$p = 0.52$	$p < 0.001$	$p = 0.001$	

accurate productions across all tested images. Whereas, Category six CMF's reproduced images consistently ranked within the top two positions for this observer. As the results showed, it was easy to identify at least one category of CMF that accurately represented the visual perception for any given observer, and some category CMFs consistently underperform for this observer. For most observers, the results of image two were different relative to those of other images. This image has a complex mixture of hues, including green, purple, and blue. The feedback from the observers was that the color sometimes matched for one color but failed to do so for others. Therefore, they have to choose based on matching one color. This result implied a significant limitation in the current CMF categorization and suggested a potential need for improvement. This analysis indicates that a suitable CMF could aid in the accuracy of image reproduction.

The consistency of results across different images is presented in Table 3. The correlations of the z-scores between the four different image scenes, as well as the average across images, were calculated along with their corresponding p-values, all of which are listed in Table 3. A higher correlation indicates greater consistency between the two compared images. The correlation results demonstrate that most observers showed consistent judgments across different images. The null hypothesis, which posits no relationship between the test data sets, was rejected for most correlations as the p-values were below an α of 0.05. Images 1 and 4, which share similar hues, were expected to have high correlations, and indeed, seven out of eight observers had correlations above 0.7. Most observers' average results were highly correlated with three or four images, except for Observer 6. Half of the observers had lower correlations between the average results and Image 2. Despite the differences in hue among the chosen images, the results indicate that consistency was well-maintained throughout the experiment.

Discussion

The consolidated results from the three experiments are presented in Table 2. Because we were looking for consistency across experiments, which would indicate that the behavior for each task corresponded to similar categorical CMFs, rank orders and correlations of the z-scores were computed. For example, for Observer 1, the top part of the table shows that Category one ranked 2nd for CMF and 1st for Fidelity, while only 5th for Preference. The correlation between CMF and Fidelity z-scores was 0.68, with $p = 0.02$, indicating a significant correlation. Correlations between CMF and Preference as well as Fidelity and Preference were lower. The table shows a consistent pattern in the ranking of both the best and worst CMF matches across the various experiments. A majority of the observers exhibited distinct best and worst matches. The table also listed the correlation between the three experiments: for all observers except 3, the correlation between CMF and Fidelity is higher than the correlation between CMF and Preference and the correlation between Fidelity and Preference. The p values of the correlations are also presented in the table. The null hypothesis is that there is no relationship between the observed phenomena. More than half of the observers demonstrated a significant relationship between the CMF and fidelity at the $\alpha = 0.05$ level. However, none of the observers' data indicated a significant relationship between CMF and preference.

Higher correlations between CMF and Fidelidy emphasize the pivotal role of the CMF in determining the fidelity of image reproduction, whereas the lower correlations for Preference are expected because preference may be affected by factors other than the CMF itself. Preference variations have been shown to be affected by many factors, such as content [17] and culture [29]. The contrast between the Fidelity and Preference results give confidence that the CMF experiment, using a fast and efficient method instead of the traditional approach to determine the CMF category, was still representative of the observer. Additionally, these results showed the perception variation among different observers.

The observers in this study belong to a younger demographic. As a result, Category ten, which is known to correspond to older observers, consistently performed worse for a majority of the observers. This indicates the notion that CMFs play an important role in the fidelity of image reproduction. Comparing two types of standard observer CMFs, the table also indicated that the CIE06 standard observer CMF (Category one) is better than the CIE 1964 CMF (Category eleven) for fidelity in image reproduction.

Comparing results from fidelity and preference experiments (Figures 5 column two and three), observers 2 and 6 both preferred the Category five CMF's image reproduction. However, it was not a good representative CMF for observers 2 and 6 in the Fidelity experiment. Similar cases could be found in other observers' results. Category one CMF was a good representative for observers 3 and 4 but turned out to be the least preferred one among all the reproduced images. However, there were still some similarities between the representative CMF and the reproduced image's preference. As for the Category ten CMF, it was the worst for most observers, and it was also one of the least preferred images. This finding further indicates that preferences are affected by more factors other than CMF.

Moreover, there were two cases where the best-fidelity CMFs for two observers closely matched the poorer-performing CMFs. This anomaly might be attributed to the lack of an appropriate representative CMF for these specific observers. Notably, the fidelity rankings for these two observers were almost identical, suggesting that further examination of the unique characteristics of these two observers is needed. In such scenarios, even though the better-performing CMFs do not offer a perfect match, the less-performing ones tend to align more closely.

Furthermore, the color gamut inherent to a display introduces another layer of complexity to image reproduction. There are instances where specific colors might not be encompassed within the display's color gamut. Such scenarios demand remapping processes, which could inadvertently affect the fidelity of the image. It is also worth noting that the present observer categories of CMFs might not be entirely flawless. This implies a need for a more refined CMFs that mirror the attributes of various color-normal observers more closely. Lastly, the hard copy and soft copy stimuli were different in size, which may cause some latent effects. That should be considered in future research.

Conclusion

Three experiments, determining the CMFs, fidelity, and preference, were conducted to investigate the connection between the CMF and both fidelity and preference in the image reproduction

pipeline.

Traditional CMF experiments require specific instruments and are very time-consuming. A method, which was aimed to quickly determine an approximate CMF function for each observer, was conducted by using a tri-pixel display along with a set of categorized observer CMFs. This approach helped associate each individual observer to CMF categories that closely resemble their unique color vision characteristics. The results demonstrated the significance of individual differences in color perception that were presented by the variation in different CMFs.

The preference experiment showed that participants' color preferences were influenced by factors beyond just CMF. Perhaps cognitive elements, such as memory color, image content, region of interest, and others. Meanwhile, the correlation of the image fidelity experiment results and the CMF experiment results demonstrates that the personalized CMFs of participants improved the accuracy of image reproduction. This finding highlights a trend between the accuracy of color representation and the alignment of the CMF.

These results demonstrate that adopting individualized CMFs in the image reproduction pipeline can significantly enhance the accuracy. This can be considered a proof of concept of a feasible way to estimate individual CMFs and then use them in the image reproduction pipeline to produce more accurate images.

References

[1] "Cie 170-2:2015 fundamental chromaticity diagram with physiological axes - part 2: Spectral luminous efficiency functions and chromaticity diagrams," *Color research and application*, vol. 41, no. 2, pp. 216–216, 2016.

[2] R. L. Alfvin and M. D. Fairchild, "Observer variability in metameric color matches using color reproduction media," *Color Research & Application*, vol. 22, no. 3, pp. 174–188, 1997.

[3] Y. Asano and M. D. Fairchild, "Categorical observers for metamerism," *Color Research & Application*, vol. 45, no. 4, pp. 576–585, 2020.

[4] Y. Asano, M. D. Fairchild, and L. Blonde, "Individual colorimetric observer model," *PloS One*, vol. 11, no. 2, pp. e0 145 671–e0 145 671, 2016.

[5] R. S. Berns, *Billmeyer and Saltzman's Principles of Color Technology*. Wiley, 2019. [Online]. Available: https://books.google.com/books?id=8GGLDwAAQBAJ

[6] ——, "Methods for characterizing crt displays," *Displays*, vol. 16, no. 4, pp. 173–182, 1996.

[7] R. Berns, "Theory and practice of dual-rgb imaging," *Rochester: Studio for Scientific Imaging and Archiving of Cultural Heritage, Rochester Institute of Technology*, 2016.

[8] T. D. Canham, D. L. Long, M. D. Fairchild, and M. Bertalmío, "Physiologically personalized color management for motion picture workflows," *SMPTE Motion Imaging Journal*, vol. 131, no. 2, pp. 8–16, 2022.

[9] T. J. Downs, O. Kuzio, and M. J. Murdoch, "Image based measurement of augmented reality displays and stimuli," in *International Colour Association (AIC) Conference 2022*, 2022, p. 62.

[10] K. J. Emery and M. A. Webster, "Individual differences and their implications for color perception," *Current opinion in behavioral sciences*, vol. 30, pp. 28–33, 2019.

[11] P. G. Engeldrum, "Psychometric scaling: a toolkit for imaging systems development," *(No Title)*, 2000.

[12] H.-U. Kim, Y. J. Koh, and C.-S. Kim, "Pienet: Personalized image enhancement network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 374–390.

[13] O. Kuzio and S. Farnand, "Led-based versus filter-based multispectral imaging methods for museum studio photography," in *Proceedings of the International Colour Association Conference*, 2021, pp. 639–644.

[14] M. J. Murdoch, "Dynamic color control in multiprimary tunable led lighting systems," *Journal of the Society for Information Display*, vol. 27, no. 9, pp. 570–580, 2019.

[15] A. D. North and M. D. Fairchild, "Measuring color-matching functions. part i," *Color Research & Application*, vol. 18, no. 3, pp. 155–162, 1993.

[16] Y. Park and M. J. Murdoch, "Efficiently evaluating the effect of color gamut and spectral bandwidth on observer metamerism in high dynamic range displays," *Journal of the Society for Information Display*, vol. 29, no. 9, pp. 704–722, 2021.

[17] E. Pei, H. Lee, E. Fedorovskaya, and S. Farnand, "Evaluation of subjective video quality of television displays," *Electronic Imaging*, vol. 36, pp. 1–10, 2024.

[18] A. Sarkar, L. Blondé, P. Le Callet, F. Autrusseau, P. Morvan, and J. Stauder, "Toward reducing observer metamerism in industrial applications: Colorimetric observer categories and observer classification." in *Color Imaging Conference*, 2010, pp. 307–313.

[19] C. Shen, "Color change: Illumination and observers," *Rochester Institute of Technology*, 2024.

[20] T. Smith and J. Guild, "The cie colorimetric standards and their use," *Transactions of the Optical Society*, vol. 33, no. 3, p. 73, 1931.

[21] N. I. Speranskaya, "Determination of spectral colour coordinates for twenty-seven normal observers," *Opt. Spectrosc.*, vol. 7, pp. 424–428, 1959.

[22] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[23] W. S. Stiles and J. M. Burch, "N.p.l. colour-matching investigation: Final report (1958)," *Opt. Acta*, vol. 6, pp. 1–26, 1959.

[24] G. M. Thurston, D. L. Hayden, P. Burrows, J. I. Clark, V. G. Taret, J. Kandel, M. Courogen, J. A. Peetermans, M. S. Bowen, D. Miller *et al.*, "Quasielastic light scattering study of the living human lens as a function of age," *Current Eye Research*, vol. 16, no. 3, pp. 197–207, 1997.

[25] M. A. Webster and D. I. MacLeod, "Factors underlying individual differences in the color matches of normal observers," *JOSA A*, vol. 5, no. 10, pp. 1722–1735, 1988.

[26] W. D. Wright, "A re-determination of the trichromatic coefficients of the spectral colours," *Transactions of the Optical Society*, vol. 30, no. 4, p. 141, 1929.

[27] Y. Yuan, "Observer metamerism quantification and visualization," *Rochester Institute of Technology*, 2021.

[28] Y. Yuan, M. J. Murdoch, and M. D. Fairchild, "A multipri-

mary lighting system for customized color stimuli," *Color Research & Application*, vol. 47, no. 1, pp. 74–91, 2022.

[29] Y. Zhu, M. R. Luo, S. Fischer, P. Bodrogi, and T. Q. Khanh, "The effectiveness of colour appearance attributes for enhancing image preference and naturalness," in *Color and Imaging Conference*, vol. 24. Society for Imaging Science and Technology, 2016, pp. 231–236.