2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897498

A DATABASE OF VISUAL COLOR DIFFERENCES OF MODERN SMARTPHONE PHOTOGRAPHY

Keshuo Xu[1†], *Zhihua Wang*[2†], *Yang Yang*[3], *Jianlei Dong*[3], *Lihao Xu*[4], *Yuming Fang*[1], *and Kede Ma*[5]

[1]School of Information Technology, Jiangxi University of Finance and Economics
[2]Shanghai AI Laboratory [3]Guangdong OPPO Mobile Telecommunications Corp., Ltd
[4]School of Digital Media and Art Design, Hangzhou Dianzi University
[5]Department of Computer Science, City University of Hong Kong

ABSTRACT

Measures for visual color differences (CDs) are pivotal in hardware and software upgrading of modern smartphone photography. Towards this goal, we construct currently the largest database for visual CDs of smartphone photography. Our database consists of $15,335$ natural images 1) captured by six latest flagship smartphones, 2) altered by Photoshop®, 3) post-processed by built-in filters of smartphones, and 4) reproduced with incorrect color profiles. Moreover, we conduct a large-scale psychophysical experiment to gather visual CDs of $30,000$ image pairs from 20 human subjects in a well-designed laboratory environment. Last, we apply our human-rated database to compare a total of 27 classical and recent CD metrics. We show that existing metrics are limited in assessing CDs of smartphone photography, and point out promising future directions of learning-based CD metrics.

***Index Terms*—** Color difference, color reproduction, smartphone photography.

1. INTRODUCTION

Nowadays, a smartphone is more of a camera than a phone. It has become the standard digital device to take pictures and record events. Moreover, it is widely acknowledged that the biggest selling point of smartphones is their photo-taking quality, which spurs the manufacturers to upgrade the hardware and software of the integrated camera systems at an accelerated pace. Among all visually significant attributes that jointly determine photo quality, color plays an increasingly important role, especially in smartphone photography.

Color is not merely a physical property associated with an object. It is a visual sensation that may be affected by luminants, viewing conditions, and the state of the eye's adaptation [1]. Arguably, the most important perceptual aspect of color in smartphone photography is *color quality* (*i.e.*, color preference), which is, however, highly subjective and culturally conditioned. Thus, it is of more practical importance to focus on color difference (CD), which is a fundamental research problem, considered by researchers dating back to Helmholtz and Schrödinger [2], Wright and Pitt [3], and David MacAdam [4]. In early days, the psychophysical experiments [5] conducted to support the development of CD formulas were mainly based on the measurements of human colorimetric tolerances using large-size uniform color patches [6]. Few of the resulting CD datasets were publicly available. Moreover, CD formulas calibrated by the collected subjective data show limited performance in predicting the CDs of natural images as perceived by the human visual system (HVS), which takes into account spatial patterns and contexts [6, 7].

To study, develop, and recommend computational methods for evaluating CDs of color images, the Commission Internationale de l'Èclairage (CIE, International Commission

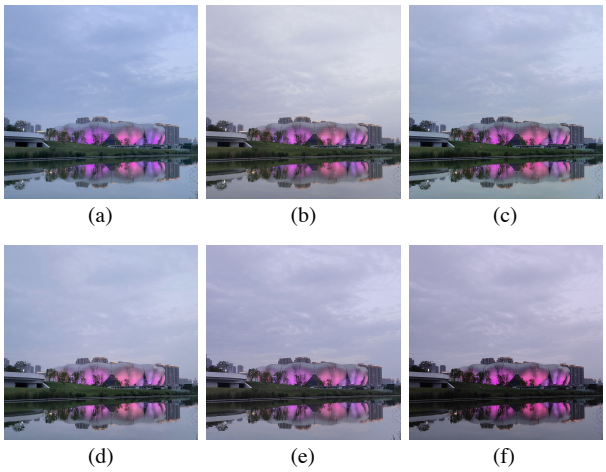

Fig. 1. Sample images captured by six flagship smartphones at same location using the night mode. (a) Apple iPhone 12 Pro. (b) HUAWEI Mate40 Pro. (c) OnePlus 7 Pro. (d) Samsung S21 Ultra. (e) OPPO Find X3 Pro. (f) Xiaomi 11 Ultra.

[†]Equal contribution

on Illumination) established the technical committee 8-02, which was closed in 2001 with a published technical report. This report lists a series of psychophysical experiments intended to calibrate CD formulas for color images by measuring the corresponding human colorimetric tolerances. Nevertheless, the numbers of reference color images in these experiments are too small to sufficiently represent the natural image manifold. Moreover, the color transforms used to induce CDs are mostly linear or quasi-linear, which are over-simplified in the age of smartphone photography. For example, Stokes [8] designed an experiment using six reference images with only systematic changes in the cathode-ray tube (CRT) display. Uroz *et al.* [9] tested four printed images with systematic and random color changes. Liu *et al.* [10] collected a CD dataset including 100 images (5 references $\times$ 20 reproductions) annotated using a categorical judgment method for the optimization of CIEDE2000 [5]. Due to the lack of large-scale visual CD datasets, it is reasonable to question the generalizability of existing CD metrics to color images captured by smartphones with great scene complexity and content diversity.

To measure the progress of CD assessment and facilitate the development of reliable CD metrics, we carry out so far the most comprehensive CD study for smartphone photography. Our contributions include:

- A large-scale dataset, including $15,335$ color images, to cover a wide range of naturally occurring situations;
- A comprehensive psychophysical experiment to collect human judgments of CDs, where we assemble $30,000$ image pairs, each of which is displayed on a carefully color-calibrated monitor and rated by a panel of at least 20 subjects in a well-controlled laboratory environment;
- A thorough performance evaluation and comparison of 27 existing CD metrics, where we find that nearly all metrics are limited in assessing CDs of smartphone photography.

2. PROPOSED DATASET

2.1. Dataset Preparation

Image Selection. We gather a total of $15,335$ color images out of $1,000$ distinct natural scenes, among which $4,002$ are captured by the authors with smartphones, 333 are downloaded from the Internet that carry Creative Commons licenses, and the remaining are color transformations of the former two. In accordance with the DXOMARK's tests[1], the natural scenes are selected to span a variety of realistic shooting scenarios, in terms of (1) *content diversity*: animal, plant, human, food, landscape, and cityscape; (2) *background complexity*: cluttered and single-color; (3) *lighting condition*: diffuse light, front light, back light, natural light in the sunrise, noon, sunset, and night; (4) *weather condition*: sunshiny, cloudy, and rainy; (5) *camera mode*: high-dynamic-range (HDR), night, etc. After that, all images are resized and cropped to 1024×1024 to combat possible compression artifacts.

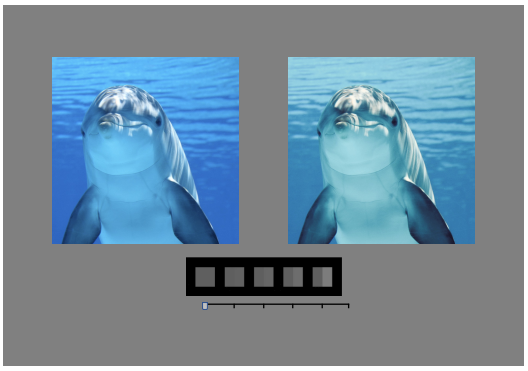

Fig. 2. The graphical user interface for subjective testing.

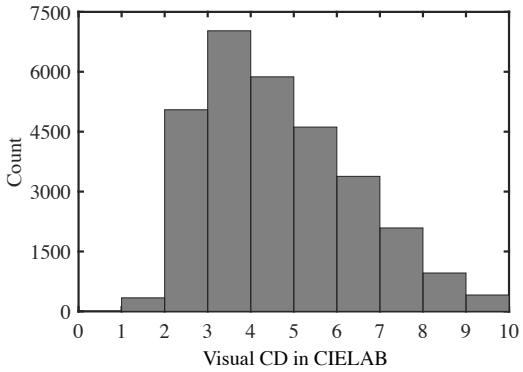

Fig. 3. Empirical distributions of $30,000$ visual CDs in the proposed dataset.

CD Generation. We create four types of CDs that are naturally occurring in smartphone photography:

- Same scene captured by different smartphones. We use six flagship smartphones, namely (a) Apple iPhone 12 Pro, (b) HUAWEI Mate40 Pro, (c) OnePlus 7 Pro, (d) Samsung S21 Ultra, (e) OPPO Find X3 Pro, and (f) Xiaomi 11 Ultra. Due to the fact that the camera system and the associated image signal processor (ISP) are proprietary, and vary among different smartphone brands, the captured pictures inevitably exhibit different color appearances, especially in night scenes (see Fig. 1). One subtlety is that different cameras may produce images of different sizes and displacements, which require cropping and alignment. We adopt a simple feature-based method[2] to estimate an affine matrix for global (and possibly non-perfect) registration.
- Same image altered by Photoshop to simulate ISP

[1] https://www.dxomark.com

[2] https://github.com/khufkens/align_images

Table 1. Min, max, median, and mean SRCC, PLCC, and STRESS between two randomized subgroups with equal size across 100 splits

Criterion	Min	Max	Median	Mean
SRCC	0.823	0.887	0.866	0.864
PLCC	0.819	0.890	0.869	0.866
STRESS	17.092	23.702	18.750	19.263

functions. Since white balance, color correction, gamma correction, and tone-mapping are four main sub-modules in the ISP relevant to color reproduction and manipulation, we synthesize these color transforms by adjusting the corresponding parameters in Photoshop, respectively.

- Same image post-processed by built-in filters of the Apple iPhone 12 Pro. We select nine filters to produce different artistic styles: *vivid*, *vivid warm*, *vivid cool*, *dramatic*, *dramatic warm*, *dramatic cool*, *mono*, *silver tone*, and *noir*[3].
- Same image reproduced with incorrect International Color Consortium (ICC) profiles. This may be the primary reason when a color management system fails to maintain the color appearance of a natural image across media devices. As an instance, an sRGB image may look over-saturated on a monitor that supports a wide color gamut, *e.g.*, DCI-P3 and Rec. 2020. We simulate two cases: sRGB images mis-display in DCI-P3 and vice versa.

In the proposed dataset, the percentages of the four types of images are 26.1%, 52.2%, 13.0%, and 6.5%, respectively. Finally, we randomly sample $20,000$ pairs of non-perfectly aligned images from the first type and $10,000$ pairs of perfectly aligned images from the remaining three types, leading to a total of $30,000$.

2.2. Psychophysical Experiment

Experimental Setup. The subjective testing environment is setup as a completely dark indoor office with no illumination and little reflection. A customized graphical user interface is devised for human data collection. As shown in Fig. 2, the background is set to be neutral gray. A pair of images with the same content but different color appearances are displayed in full resolution, whose CD is rated with reference to five pairs of grayscale samples. The lightness differences of the five grayscale pairs in the CIELAB unit (ΔE^*_{ab}) are around 0, 1.7, 3.4, 6.8, and 13.6, respectively. A scale-and-slider applet is located at the bottom to collect continuous scores. The viewing distance is fixed to one meter. Ten male and ten female observers, who have normal color vision and normal or corrected-to-normal visual acuity, participate in the psychophysical experiment.

Display Characterization. Two EIZO CG319X 31.1" LCD monitors are adopted in the experiment. The display peak white is set to be 100 cd/m^2 with a correlated color temperature of $6,500$ degrees of Kelvin. The color performance of the display is thoroughly checked and calibrated. We use CIE recommended gamma-offset-gain display model for characterization, *i.e.*, modeling the non-linear relationship between the digital input and the luminance of each channel in RGB color space. The gain, offset, and gamma of each channel in RGB are measured by a tele-spectroradiometer - JETI Specbos 1211, whose accuracy is within 2% when measuring Illuminant A of 100 cd/m^2, under the assumption of the CIE 1964 standard colorimetric observer (*i.e.*, 10° observer). The obtained model is tested to have a performance of 0.56 ΔE^*_{ab} based on the Macbeth ColorChecker Chart, indicating that our display is suitable for color-related vision experiments.

2.3. Visual CDs

Outlier Detection and Subject Rejection. For ease of analysis, the raw subjective score (with reference to the grayscale grade) should be transformed to the CIELAB unit, *i.e.*, ΔE^*_{ab}, which is commonly referred to as the visual CD. Next, we ensure the reliability of the collected visual CDs. For the visual CDs that lie out of 3 standard deviations are identified as outliers and subsequently discarded. Subjects with an outlier rate $\geq 5\%$ are considered invalid and are rejected. After data purification, we find that all subjects are valid, and 1.09% of the ratings are detected as outliers. The mean of valid visual CDs of each image pair is considered as the ground-truth.

Analysis. Fig. 3 plots the histogram of visual CDs for $30,000$ image pairs in the proposed dataset, which is well fitted by a unimodal distribution with the mode around 2. To verify the reliability of the collected CDs, we randomly split the subjects into two subgroups of equal size, and compute the Spearman's rank correlation coefficient (SRCC), the Pearson linear correlation coefficient (PLCC) and the standardized residual sum of squares (STRESS) [35] between their respective mean visual CDs. We repeat this procedure for 100 times, and show the results in Table 1, where high consistency between two subgroups have been observed.

3. APPLICATION OF THE PROPOSED DATASET

We apply the proposed dataset to compare 27 state-of-the-art computational methods that are used for CD assessment. The results in Table 2 are organized in terms of color-patch-oriented and color-image-oriented methods, respectively. According to the algorithm design objectives, the color-image-oriented methods are further categorized into CD metrics [5–

[3] `https://backlightblog.com/iphone-filters-effects` shows the detailed descriptions of the nine artistic styles.

Table 2. SRCC, PLCC and STRESS between predicted CDs and visual CDs in our dataset. Top section lists the representative CD formulas developed from color patch data. Second section contains CD metrics intended for natural color images

Method	Color space	Perfectly aligned pairs			Non-perfectly aligned pairs			All		
		STRESS	PLCC	SRCC	STRESS	PLCC	SRCC	STRESS	PLCC	SRCC
CIELAB [11]	CIELAB	31.242	0.794	0.776	29.859	0.690	0.580	31.832	0.716	0.668
CMC [12]	CIELAB	33.997	0.790	0.788	34.202	0.594	0.493	35.796	0.666	0.634
CIE94 [13]	CIELAB	34.670	0.793	0.774	29.952	0.699	0.574	34.213	0.712	0.656
CIEDE2000 [5]	CIELAB	29.837	0.831	0.824	30.406	0.672	0.564	31.283	0.728	0.689
Huertas06 [14]	OSA-UCS	36.860	0.680	0.687	34.463	0.560	0.426	36.608	0.576	0.561
S-CIELAB[1][7]	CIELAB	29.952	0.829	0.822	31.834	0.634	0.524	32.608	0.702	0.660
Imai01 [15]	CIELAB	60.517	0.691	0.703	48.010	0.530	0.538	57.335	0.603	0.617
Hong06 [6]	CIELAB	60.602	0.808	0.813	58.252	0.542	0.457	61.849	0.651	0.632
Chou07 [16]	CIELAB	50.908	0.794	0.786	36.013	0.615	0.459	49.488	0.616	0.558
Simone09 [17]	OSA-UCS	36.109	0.688	0.696	35.055	0.543	0.399	36.757	0.565	0.545
Pedersen12 [18]	CIELAB	60.365	0.810	0.814	58.493	0.479	0.400	63.208	0.609	0.598
Lissner12 [19]	CIELAB	36.437	0.608	0.620	40.232	0.332	0.242	41.189	0.428	0.419
Toet03 [20]	$\ell\alpha\beta$	34.451	0.400	0.391	39.031	0.149	0.050	36.008	0.229	0.160
Pinson04 [21]	YC_bC_r	51.524	0.310	0.280	59.811	0.075	0.034	59.102	0.220	0.150
SSIM [22]	Gray scale	38.911	0.584	0.544	52.859	0.054	0.010	47.971	0.308	0.164
Lee05 [23]	CIELAB	58.390	0.735	0.742	56.478	0.639	0.644	57.833	0.701	0.717
Ouni08[1][24]	CIELAB	29.839	0.830	0.824	30.414	0.672	0.564	31.286	0.728	0.689
Yu09 [25]	HSI	69.463	0.294	0.315	68.121	0.267	0.233	69.140	0.273	0.295
Ponomarenko11 [26]	YC_bC_r	49.861	0.530	0.532	47.697	0.121	0.081	52.484	0.298	0.222
Gao13 [27]	OCC	63.105	0.245	0.210	60.436	0.361	0.250	62.603	0.280	0.229
Lee14 [28]	CIELAB	45.668	0.579	0.584	39.579	0.290	0.240	54.529	0.361	0.271
VSI [29]	LMN	34.731	0.661	0.666	39.354	0.141	0.095	36.258	0.243	0.264
Jaramillo19 [30]	YC_bC_r	43.493	0.512	0.501	50.120	0.054	0.011	68.440	0.297	0.160
Butteraugli [31]	RGB	42.690	0.612	0.589	48.764	0.245	0.178	54.801	0.362	0.323
ꟻLIP [32]	CIELAB	29.365	0.743	0.714	27.565	0.730	0.638	29.197	0.716	0.663
PieAPP [33]	RGB	41.546	0.502	0.512	39.625	0.483	0.410	41.896	0.467	0.451
LPIPS [34]	RGB	47.162	0.670	0.679	40.307	0.253	0.233	66.604	0.407	0.259

[1] The spatial extension of CIEDE2000.

7, 11–15, 17–20, 23, 24, 28, 30], general-purpose image quality models [21, 22, 25–27, 29], and just noticeable difference (JND) measures [16, 19, 31–34]. We find surprisingly that the color-patch-oriented methods perform favorably on perfectly aligned pairs. Compared to CIEDE2000, the performance of its two spatial extensions, S-CIELAB and Ouni08 even drops slightly, indicating that simple spatial filtering considering the contrast sensitivity of the HVS seems ineffective for CD assessment. Overall, the CD assessment performance is better on perfectly aligned pairs than that on non-perfectly aligned pairs, captured by different smartphones. From the experimental results, we are able to conclude that existing CD metrics are limited in assessing the CDs of smartphone photography, especially when there is misalignment between the two images, imperceptible to the human eye, though.

4. CONCLUSION

In this paper, we look at a challenging and long overlooked problem of CD assessment for smartphone photography. We build a large-scale CD dataset consisting of $30,000$ image pairs that cover most naturally occurring situations in smartphone photography. Each image pair receives at least 20 human ratings with verified reliability. Based on this dataset, we make a comprehensive comparison of 27 state-of-the-art CD methods. Experiment results indicate that none of the CD methods is able to achieve high correlations and low STRESS. We hope our newly established dataset becomes a new benchmark for the development of CD metrics. A promising future direction is to construct a learning-based and end-to-end optimized CD method based on the proposed dataset, which generalizes CIELAB-based metrics, and delivers superior CD assessment performance in the presence of misalignment.

5. REFERENCES

[1] Commission Internationale de l'Èclairage, *Colorimetry*. CIE Publication, 2004.

[2] R. Kühni, "Historical development of color space and color difference formulas," *Color Space and Its Divisions*, 2003.

[3] W. Wright and F. Pitt, "Hue-discrimination in normal colour-vision," *Proc Phys Soc*, vol. 46, no. 3, pp. 459–473, 1934.

[4] D. MacAdam, "Visual sensitivities to color differences in daylight," *JOSA*, vol. 32, no. 5, pp. 247–274, 1942.

[5] M. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Res Appl*, vol. 26, no. 5, pp. 340–350, 2001.

[6] G. Hong and M. Luo, "New algorithm for calculating perceived colour difference of images," *Imaging Sci J*, vol. 54, no. 2, pp. 86–91, 2006.

[7] X. Zhang and B. Wandell, "A spatial extension of CIELAB for digital color-image reproduction," *J Soc Inf Display*, vol. 5, no. 1, pp. 61–63, 1997.

[8] M. Stokes, "Colorimetric tolerances of digital images," Master's thesis, Rochester Institute of Technology, 1991.

[9] J. Uroz, M. Luo, and J. Morovic, *Colour Image Science: Exploiting Digital Media*. John Wiley & Sons, Ltd, 2002.

[10] H. Liu, M. Huang, G. Cui, *et al.*, "Color-difference evaluation for digital images using a categorical judgment method," *JOSA A*, vol. 30, no. 4, pp. 616–626, 2013.

[11] A. Robertson, "The CIE 1976 color-difference formulae," *Color Res Appl*, vol. 2, no. 1, pp. 7–11, 1977.

[12] British Standards Institution, "Method for calculation of small colour differences," *American National Standards Institute*, 1998.

[13] R. McDonald and K. Smith, "CIE94-A new colour-difference formula," *J Soc Dyers Colour*, vol. 111, no. 12, pp. 376–379, 1995.

[14] R. Huertas, M. Melgosa, and C. Oleari, "Performance of a color-difference formula based on OSA-UCS space using small-medium color differences," *JOSA A*, vol. 23, no. 9, pp. 2077–2084, 2006.

[15] F. Imai, N. Tsumura, and Y. Miyake, "Perceptual color difference metric for complex images based on mahalanobis distance," *J Electron Imaging*, vol. 10, no. 2, pp. 385–393, 2001.

[16] C. Chou and K. Liu, "A fidelity metric for assessing visual quality of color images," in *ICCCN*, pp. 1154–1159, 2007.

[17] G. Simone, C. Oleari, and I. Farup, "An alternative color difference formula for computing image difference," in *GCIS*, no. 4, pp. 8–11, 2009.

[18] M. Pedersen and J. Hardeberg, "A new spatial filtering based image difference metric based on hue angle weighting," *J Imaging Sci Techn*, vol. 56, no. 5, pp. 1–12, 2012.

[19] I. Lissner, J. Preiss, P. Urban, *et al.*, "Image-difference prediction: From grayscale to color," *IEEE TIP*, vol. 22, no. 2, pp. 435–446, 2012.

[20] A. Toet and M. Lucassen, "A new universal colour image fidelity metric," *Displays*, vol. 24, no. 4-5, pp. 197–207, 2003.

[21] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE TOB*, vol. 50, no. 3, pp. 312–322, 2004.

[22] Z. Wang, A. Bovik, H. Sheikh, *et al.*, "Image quality assessment: From error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[23] S. Lee, J. Xin, and S. Westland, "Evaluation of image similarity by histogram intersection," *Color Res Appl*, vol. 30, no. 4, pp. 265–274, 2005.

[24] S. Ouni, E. Zagrouba, M. Chambah, *et al.*, "A new spatial colour metric for perceptual comparison," in *ICESECI*, pp. 413–428, 2008.

[25] M. Yu, H. Liu, Y. Guo, *et al.*, "A method for reduced-reference color image quality assessment," in *ICISP*, pp. 1–5, 2009.

[26] N. Ponomarenko, O. Ieremeiev, V. Lukin, *et al.*, "Modified image visual quality metrics for contrast change and mean shift accounting," in *CADSM*, pp. 305–311, 2011.

[27] C. Gao, K. Panetta, and S. Agaian, "No reference color image quality measures," in *ICC*, pp. 243–248, 2013.

[28] D. Lee and K. Plataniotis, "Towards a novel perceptual color difference metric using circular processing of hue components," in *ICASSP*, pp. 166–170, 2014.

[29] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE TIP*, vol. 23, no. 10, pp. 4270–4281, 2014.

[30] B. Ortiz-Jaramillo, A. Kumcu, L. Platisa, *et al.*, "Evaluation of color differences in natural scene color images," *SPIC*, vol. 71, pp. 128–137, 2019.

[31] J. Alakuijala, R. Obryk, O. Stoliarchuk, *et al.*, "Guetzli: Perceptually guided JPEG encoder," *arXiv preprint arXiv:1703.04421*, 2017.

[32] P. Andersson, J. Nilsson, T. Akenine-Möller, *et al.*, "ꟻLIP: A difference evaluator for alternating images," *ACM SIGGRAPH*, vol. 3, no. 2, pp. 1–23, 2020.

[33] E. Prashnani, H. Cai, Y. Mostofi, *et al.*, "PieAPP: Perceptual image-error assessment through pairwise preference," in *CVPR*, pp. 1808–1817, 2018.

[34] R. Zhang, P. Isola, A. A. Efros, *et al.*, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, pp. 586–595, 2018.

[35] P. A. Garcia, R. Huertas, M. Melgosa, *et al.*, "Measurement of the relationship between perceived and computed color differences," *JOSA A*, vol. 24, no. 7, pp. 1823–1829, 2007.