Rehabilitating the ColorChecker Dataset for Illuminant Estimation

Ghalia Hemrit[1], *Graham D. Finlayson*[1], *Arjan Gijsenij*[2], *Peter Gehler*[3], *Simone Bianco*[4], *Brian Funt*[5], *Mark Drew*[5] *and Lilong Shi*[6]

[1] *School of Computing Sciences, University of East Anglia; Norwich, United Kingdom*

[2] *AkzoNobel; Amsterdam, The Netherlands*

[3] *Amazon; Tübingen, Germany*

[4] *Imaging and Vision Laboratory, University of Milan-Bicocca; Milan, Italy*

[5] *School of Computing Science, Simon Fraser University; Vancouver, Canada*

[6] *Samsung Semiconductor Inc; Pasadena, USA*

Abstract

In a previous work, it was shown that there is a curious problem with the benchmark ColorChecker dataset for illuminant estimation. To wit, this dataset has at least 3 different sets of ground-truths. Typically, for a single algorithm a single ground-truth is used. But then different algorithms, whose performance is measured with respect to different ground-truths, are compared against each other and then ranked. This makes no sense. We show in this paper that there are also errors in how each ground-truth set was calculated. As a result, all performance rankings based on the ColorChecker dataset – and there are scores of these – are inaccurate.

In this paper, we re-generate a new 'recommended' ground-truth set based on the calculation methodology described by Shi and Funt. We then review the performance evaluation of a range of illuminant estimation algorithms. Compared with the legacy ground-truths, we find that the difference in how algorithms perform can be large, with many local rankings of algorithms being reversed.

Finally, we draw the readers attention to our new 'open' data repository which, we hope, will allow the ColorChecker set to be rehabilitated and once again become a useful benchmark for illuminant estimation algorithms.

1. Introduction

The ColorChecker dataset was introduced by Gehler et al. in 2008 [1]. It has 568 images of various daily and ordinary tourist scenes (see Figure 1) mainly taken in Cambridge with two popular cameras, the Canon 1D and the Canon 5D. The ColorChecker dataset is probably the most widely used dataset in evaluating the performance of algorithms for illuminant estimation.

The goal of most illuminant estimation algorithms is to infer the chromaticity of the light. The reader will notice that each image in Figure 1 has a Macbeth ColorChecker placed in the scene. The RGB from the brightest non-saturated achromatic patch (see last row of color chart) is the correct answer – or ground-truth – for illuminant estimation, according to the calculation methodology described by Shi and Funt [2]. In fact, most of the recent work in illuminant estimation uses a linear version of the dataset, with linear images reprocessed by Shi and Funt [2] from Gehler's original raw images [1].

Naturally, the performance of illuminant estimation algorithms – which adopt a range of strategies to infer the RGB of the illuminant – is determined by how well they predict the ground-truth illuminant color. Because we cannot distinguish between

a bright scene dimly lit and the converse (because the intensity of the illuminant is not recovered), the measure which is used to quantify the accuracy of the estimation is the angular error. Given a set of angular errors, various statistical summaries are used to summarise performance. These include the mean, median or 95% quantile angular error. Given a set of algorithms and their summary performance statistics, it is natural to rank all the algorithms (according to the statistic), and then to conclude that algorithm A is better than B which is better than C (e.g. if their respective means are in an ascending order).

While there are many image sets (with respect to which, algorithms might be evaluated in performance and ranked), the ColorChecker dataset is the most widely used. Indeed, perhaps the most comprehensive survey of algorithms performance was carried out by Gijsenij et al. [3]. The results reported there are often quoted in more recent papers with the paper now cited ~400 times (April 13, 2018). The algorithms considered in this survey also form the basis of the color constancy evaluation site (color-constancy.com).
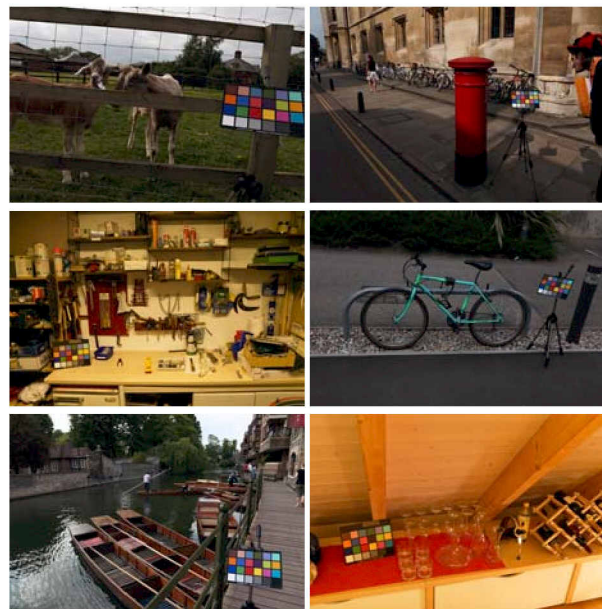

Figure 1: Images from the Macbeth ColorChecker dataset; here the images are the camera pipeline outputs.

Of course, this evaluation process makes perfect sense, *in*

principle. In practice, we observed in [4] a surprising flaw in the methodology. Namely, we discovered that there are 3 different sets of ground-truths, one ground-truth – **SFU** – on www.cs.sfu.ca/~colour/data/shi_gehler/ [2] and two ground-truths – **Gt1** and **Gt2** – on colorconstancy.com. Unfortunately, these 3 ground-truths are used in a haphazard way. Specifically, an author will adopt one of the three ground-truths, then will evaluate the algorithm and, say, calculate the median angular error. In the next step, this median is compared against other algorithms' median errors. But, the medians for these competitor algorithms have been calculated with respect to **all 3** ground-truth correct answers. This makes no sense, especially since [4] demonstrated that using different sets of ground-truths would drastically affect the ranking of algorithms. At the time of writing this paper, there is no definitive ranking of illuminant estimation algorithms (for the ColorChecker dataset).

Here, we seek to explain why we find ourselves in this multiple ground-truth world. We then go on to make a new 'recommended' ground-truth (**REC**) for the community. Then, we present the results evaluating algorithms using this recommended ground-truth and compare our results to the benchmark evaluation and rankings. The rankings relative to the new recommended ground-truth reveal for the first time the actual pecking order in illuminant estimation for the ColorChecker dataset.

In Section 2, we discuss how we compute the new **REC** ground-truth and explain why this ground-truth differs from the other 3 in the literature. In section 3, we present some analysis of the relative performance of different algorithms using **REC** compared to the legacy data. The paper ends with a short conclusion.

2. Derivation of the RECommended Ground-truth

We adopt directly the methodology introduced by Shi and Funt [2] to re-process the raw images and re-calculate the ground-truth set of illuminants of the ColorChecker dataset. However, we have used our own code and in the interests of transparency, we will make our code accessible online [5].
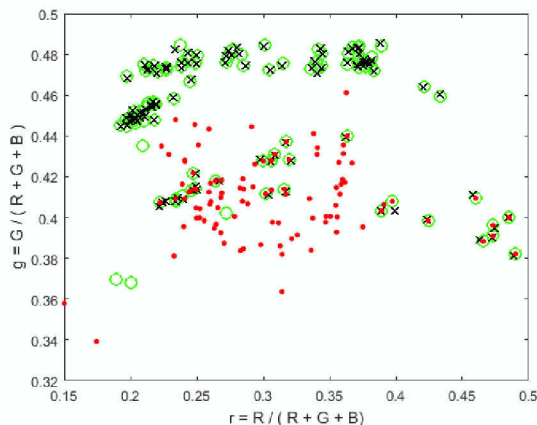

Figure 2: REC ground-truth chromaticities are plotted as black crosses. The green circles and red dots respectively denote the SFU and Gt1 ground-truths.

Readers who 'click through' to this code and the data will find some additional information. Specifically, we explain that our data repository is 'open'. We set forth a mechanism for the community to provide further suggested modifications to this ground-truth dataset (if necessary). In the short term, we will monitor any suggestions that the community makes and incorporate these suggestions – if they are significant – into the data. This would lead to a new recommended ground-truth. This said, we have been quite careful in our calculations so it is our hope that our new **REC** ground-truth will stand the test of time. If a modification is suggested, then both the current **REC** and the updated version will be retained and labelled with clear time stamps.

Let us outline the **REC** ground-truth calculation. We re-processed Gehler's raw images to obtain linear demosaiced images using *dcraw* [6]. Every image has a ColorChecker in the scene. This color chart provides a reference for measuring the ground-truth illuminant. In each image, the ColorChecker chart is selected and the median RGB from the brightest achromatic patch (ranked by average of the selected squares with no digital count>3300 [each image is in 12 bits]) defines a ground-truth illuminant. The main steps of the color chart processing are presented in Figure 3.

One point to highlight is the importance of using the same patch or patches when defining the R, G and B of the light color. We found that this property was not properly enforced in the calculation of the **SFU** ground-truth. In fact, we found 3 images where saturated color channels/patches were not correctly identified which resulted in having the ground-truth R, G and B values not taken from the same patch, in these three cases.
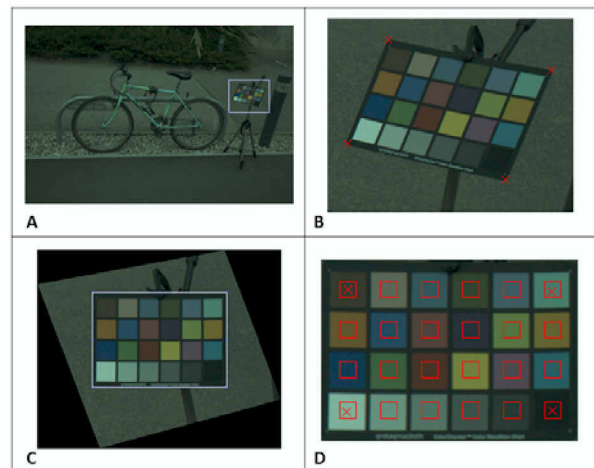

Figure 3: The 4 main steps of the color chart processing. A) we select roughly the chart area in the image. B) we select the 4 corners of the chart. C) the chart image is geometrically transformed to be in front of the camera, then we select more precisely its contour. D) we select the centers of the 4 corners patches as well as one square of interest in one patch, this square selection is then automatically replicated over all patches. The medians per channel of the achromatic red-square regions are calculated.

An important detail about the ColorChecker dataset is the 'black level'. This is zero for Canon 1D images but is 129 for the Canon 5D. One important contribution of the Shi-Funt calculation methodology is the subtraction of the camera black level from the ground-truth. The offsets were estimated from the minimum

pixels across the whole dataset [2].

Let us discuss the ground-truths. The ground-truth set we call **SFU** has the 568 estimates of the RGB lights colors and it appears currently on the SFU site. On colorconstancy.com, there are two additional ground-truths, **Gt1** and **Gt2**. **Gt1** was calculated and put online by Gijsenij in 2011. It refers to the same data as **SFU** but without the black level being subtracted. **Gt1** was calculated based on the color chart patches RGB data which is available on the SFU site [2] but without subtracting the offset. A little investigation on our part led us to the following explanation: on the *current* site [2], it is stated that the black level needs to be subtracted before using the data: "Note that for most applications . . . the black level offset will still need to be subtracted from the original images". When looking at the same web-page using the waybackmachine.org site – that allows recovering web-pages from many years ago – the previous statement is not present and instead we can read "[the processing] also takes into account that the Canon 5D has a black level of 129, which we subtract." We spoke to a number of researchers in the field and asked them to read the current and past web-pages. Everyone agreed that on reading the past instructions they could easily have made the same 'mistake' (and not subtracted the black level). Gijsenij believes the web-page instructions are a plausible explanation of why he did not subtract the black point.

Researchers who have been in the field a long time have been using **Gt1** but more recent workers are using **SFU**. Potentially, all of these researchers *think* that they are using the same ground-truth. But in fact, the data are very different.

The third legacy ground-truth dataset we call **Gt2**. It is very similar to **Gt1** (and so, also, different from **SFU**). The reason for the difference between **Gt1** and **Gt2** is explained by Bianco [22]:

Table 1: Ranking of 23 algorithms in terms of median recovery error for REC vs SFU vs Gt1; the Minkowski norm p and the smoothing value σ are the optimal parameters.

Algorithm	REC		SFU		Gt1	
	median	rank	median	rank	median	rank
Fast Fourier Color Constancy (model Q) [7]	1.14°	1	1.13°	1	3.82°	15
1st order Grey-Edge (p=1, σ=6) [8]	3.09°	2	3.08°	2	4.52°	18
Edge-based Gamut (σ=4) [9]	3.27°	3	3.26°	3	5.04°	19
2nd order Grey-Edge (p=1, σ=1) [8]	3.57°	4	3.59°	4	4.44°	17
Bayesian [1][10]	3.85°	5	3.87°	5	3.46°	14
Deep Color Constancy using CNNs [11]	3.94°	6	3.93°	6	1.99°	1
Exemplar-based Colour Constancy [12]	4.36°	7	4.37°	7	2.27°	2
Pixel-based Gamut (σ=4) [9]	4.41°	8	4.44°	8	2.33°	3
Intersection-based Gamut (σ=4) [9]	4.41°	9	4.44°	9	2.34°	4
Bottom-up + Top-down [13]	4.54°	10	4.56°	10	2.47°	5
Top-Down [13]	4.60°	11	4.63°	11	2.63°	7
Using Natural Image Statistics [14]	4.70°	12	4.72°	12	3.13°	11
Heavy Tailed-based Spatial Correlations [15]	4.76°	13	4.81°	13	2.96°	9
Bottom-Up [13]	4.90°	14	4.90°	14	2.56°	6
CART-based Selection [16]	5.09°	15	5.09°	15	3.35°	12
Alex-Net + SVR using CNNs [11]	5.30°	16	5.30°	16	3.11°	10
CART-based Combination [16]	5.52°	17	5.52°	17	2.91°	8
General Grey-World (p=9, σ=9) [8]	5.95°	18	5.95°	18	3.46°	13
White-Patch [17]	6.74°	19	6.74°	19	5.68°	20
Shades-of-Grey (p=4) [18]	6.83°	20	6.85°	20	4.01°	16
Regression (SVR) [19]	9.60°	21	9.62°	21	6.73°	22
Grey-World [20]	9.97°	22	9.99°	22	6.29°	21
Inverse-Intensity Chromaticity Space [21]	11.61°	23	11.58°	23	11.52°	23

"we noticed that for some images the Macbeth ColorChecker coordinates (both the bounding box and the corners of each patch) were wrong and thus the illuminant ground-truth was wrong." In Figure 2, for a subset of the 568 images, we compare the chromaticity distributions of our new **REC** ground-truth, **SFU** and **Gt1** (**Gt2** is not shown as it is almost the same as **Gt1**). The reader will notice that the *current* **SFU** ground-truth is close to our newly calculated **REC**ommended ground-truth except for a few points. These look to be set apart from the rest of the data i.e. they appear to be outliers, in some sense (note that, in Figure 2, the 3 green circles to the left of the plot that do not overlap with the black crosses). These are data points in **SFU** that are different from **REC**.

We posit that the outliers are due to the problems in calculating correctly the bounding boxes (Bianco's observation) and to our own discovery that the white point was on occasion drawn from different achromatic patches (for R versus G versus B). Despite this, the preponderance of the data is in, more or less, precise alignment. In contrast, the points in **Gt1** are far from **REC**.

Table 2: Ranking of 23 algorithms in terms of median reproduction error [23] for REC vs SFU vs Gt1; the Minkowski norm p and the smoothing value σ are the optimal parameters.

Algorithm	REC		SFU		Gt1	
	median	rank	median	rank	median	rank
Fast Fourier Color Constancy (model Q) [7]	1.43°	1	1.45°	1	4.75°	16
1st order Grey-Edge (p=1, σ=9) [8]	3.74°	2	3.77°	2	4.90°	18
2nd order Grey-Edge (p=1, σ=1) [8]	4.54°	3	4.59°	4	4.76°	17
Edge-based Gamut (σ=3) [9]	4.56°	4	4.54°	3	5.88°	19
Bayesian [1][10]	4.60°	5	4.62°	5	3.92°	13
Deep Color Constancy using CNNs [11]	4.72°	6	4.75°	6	2.24°	1
Exemplar-based Colour Constancy [12]	5.17°	7	5.22°	9	2.64°	2
Pixel-based Gamut (σ=4) [9]	5.20°	8	5.21°	7	2.73°	3
Intersection-based Gamut (σ=4) [9]	5.20°	9	5.21°	8	2.74°	4
Bottom-up + Top-down [13]	5.36°	10	5.35°	10	2.75°	5
Top-Down [13]	5.41°	11	5.43°	11	3.11°	7
Using Natural Image Statistics [14]	5.46°	12	5.47°	12	3.55°	10
Heavy Tailed-based Spatial Correlations [15]	5.83°	13	5.88°	13	3.48°	8
Bottom-Up [13]	5.85°	14	5.89°	14	2.98°	6
CART-based Selection [16]	6.08°	15	6.03°	15	3.90°	12
Alex-Net + SVR using CNNs [11]	6.34°	16	6.38°	16	3.66°	11
CART-based Combination [16]	6.56°	17	6.55°	17	3.48°	9
General Grey-World (p=9, σ=9) [8]	6.66°	18	6.65°	18	3.98°	14
Shades-of-Grey (p=4) [18]	7.55°	19	7.56°	19	4.43°	15
White-Patch [17]	8.04°	20	8.03°	20	6.48°	20
Regression (SVR) [19]	10.31°	21	10.39°	21	7.42°	22
Grey-World [20]	10.63°	22	10.68°	22	6.81°	21
Inverse-Intensity Chromaticity Space [21]	12.81°	23	12.77°	23	12.77°	23

3. Evaluating the Performance of Illuminant Estimation Algorithms

In what follows, we provide performance evaluation and ranking of 23 illuminant estimations for the **REC**ommended and the legacy **SFU** and **Gt1** ground-truths. In Table 1, the 23 algorithms are ranked according to the median recovery angular error (i.e. the conventional error measure, which gives the angle between the estimated illuminant RGB vector and the ground-truth illuminant RGB). The ranking is the same for **REC** and **SFU**, although the median errors are slightly different. This is expected as (recall Figure 2) the number of differences between **REC** and **SFU** is small (due to errors in the calculation). However, the rank-

ing of algorithms according to the **Gt1** dataset is markedly different (due as opposed to a different methodology of calculation).

The five best algorithms with **Gt1** are not in the top 5 according to **REC**. Vice versa, the 5 best algorithms according to **REC** are among the worst-performing algorithms with **Gt1**. The algoithms Edge-based Gamut Mapping [9], 2nd order Grey-Edge [8] and Bayesian [1][10] are, for example, in reverse order. Fast Fourier Color Constancy [7] – which is in significant part built on top of a machine learning algorithm – is best according to the **REC**ommended ground-truth but is the 15th best on **Gt1**. This may not be surprising as this algorithm was trained on **SFU**. Deep Color Constancy using CNNs [11] is 6th based on **REC** and top ranked based on **Gt1**. Again, this is not surprising since this algorithm was trained on **Gt2** (which, we recall, is very similar to **Gt1** but very different from **SFU** and **REC**).

In Table 2, we consider the ranking of the 23 illuminant estimation algorithms in terms of median reproduction angular error [23]. Reproduction error is an angle-type error that evaluates how well a white surface is reproduced. Since image reproduction is the goal of most illuminant estimation algorithms, reproduction error provides a more useful measure of algorithm performance. This time, results with **REC** and **SFU** locally differ. Once again, the results with **Gt1** are significantly different. Finally notice that the ranks for the same ground-truth but recovery vs reproduction angular error results in different rankings.

In Tables 1 and 2, we do not include the results for **Gt2** because they are very comparable to **Gt1**, however we invite the reader to consult our previous work on this topic [4]. A more complete survey is accessible on colorconstancy.com.

Conclusion

Illuminant estimation algorithms have been evaluated and compared on the benchmark ColorChecker dataset with at least three different ground-truths, with one of the three being very different to the other two. In addition, we found that all three of these sets of ground-truths were inaccurately or incorrectly calculated in the sense that small errors were made. The problem of multiple ground-truths and calculation errors has led to misleading results in the performance evaluation of illuminant estimation algorithms.

In this paper we have introduced a new **REC**ommended ground-truth for this dataset which we hope rehabilitates the ColorChecker dataset. Broadly, we followed the methodology set forth by Shi and Funt but using our own code and we corrected a few errors made (e.g. those reported in [22]). We re-evaluated all the algorithms on the widely used comparison site for illuminant estimation algorithms, colorconstancy.com. We invite the community to refer to what we hope is a more definitive comparison in future research.

Acknowledgments

References

[1] P. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian Color Constancy Revisited," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, 2008.

[2] L. Shi and B. Funt, "Re-processed Version of the Gehler Color Constancy Dataset of 568 Images." www.cs.sfu.ca/~colour/data/shi_gehler/, [accessed in March-2018].

[3] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Computational Color Constancy: Survey and Experiments," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2475–2489, 2011.

[4] G. D. Finlayson, G. Hemrit, A. Gijsenij, and P. Gehler, "A Curious Problem with using the Colour Checker Dataset for Illuminant Estimation," in *Color Imaging Conf.*, pp. 64–69, 2017.

[5] A. Gijsenij and T. Gevers, "Datasets and Results per Datasets." www.colorconstancy.com, [first updated in March-2018].

[6] D. Coffin, "Decoding raw digital photos in Linux, download Dcraw 9.27." www.cybercom.net/~dcoffin/dcraw/, [accessed in January-2018].

[7] J. T. Barron and Y.-T. Tsai, "Fast Fourier Color Constancy," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[8] J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-Based Color Constancy," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2207–2214, 2007.

[9] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Generalized Gamut Mapping using Image Derivative Structures for Color Constancy," *Int. J. Comput. Vis.*, vol. 86, no. 2, pp. 127–139, 2010.

[10] C. Rosenberg, T. Minka, and A. Ladsariya, "Bayesian Color Constancy with Non-Gaussian Models," *Adv. Neural Inf. Process. Syst. MA MIT Press*, vol. 16, 2003.

[11] S. Bianco, C. Cusano, and R. Schettini, "Color Constancy using CNNs," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 81–89, 2015.

[12] M. S. Drew and H. R. V. Joze, "Exemplar-Based Colour Constancy," in *Br. Mach. Vis. Conf.*, vol. 26, pp. 1–12, 2012.

[13] J. Van De Weijer, C. Schmid, and J. Verbeek, "Using High-Level Visual Information for Color Constancy," in *IEEE Int. Conf. Comput. Vis.*, pp. 1–8, 2007.

[14] A. Gijsenij and T. Gevers, "Color Constancy using Natural Image Statistics and Scene Semantics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 687–698, 2011.

[15] A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color Constancy with Spatio-Spectral Statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1509–1519, 2012.

[16] S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Automatic Color Constancy Algorithm Selection and Combination," *Pattern Recognit.*, vol. 43, no. 3, pp. 695–705, 2010.

[17] E. H. Land and J. J. McCann, "Lightness and Retinex Theory," *J. Opt. Soc. Am.*, vol. 61, no. 1, pp. 1–11, 1971.

[18] G. D. Finlayson and E. Trezzi, "Shades of Gray and Colour Constancy," in *Color Imaging Conf.*, no. 1, pp. 37–41, 2004.

[19] B. Funt and W. Xiong, "Estimating Illumination Chromaticity via Support Vector Regression," in *Color Imaging Conf.*, pp. 47–52, 2004.

[20] G. Buchsbaum, "A Spatial Processor Model for Object Colour Perception," *J. Franklin Inst.*, vol. 310, no. 1, pp. 1–26, 1980.

[21] R. Tan, K. Nishino, and K. Ikeuchi, "Illumination Chromaticity Estimation using Inverse-Intensity Chromaticity Space," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 673–680, 2003.

[22] S. Bianco, "Personal Communication." [January 12, 2018].

[23] G. D. Finlayson, R. Zakizadeh, and A. Gijsenij, "The Reproduction Angular Error for Evaluating the Performance of Illuminant Estimation Algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1482 – 1488, 2016.