Journal of Physics: Conference Series **3128** (2025) 012014 doi:10.1088/1742-6596/3128/1/012014

On the fair use of the ColorChecker dataset for illuminant estimation

Marco Buzzelli[1], Graham Finlayson[2], Arjan Gijsenij[3], Peter Gehler[4], Mark Drew[5], Lilong Shi[4], Luca Cogo[1] and Simone Bianco[1]

[1]University of Milano - Bicocca, Italy

[2]University of East Anglia, United Kingdom

[3]AkzoNobel, Netherlands

[4]Independent researcher

[5]Simon Fraser University, Canada

E-mail: simone.bianco@unimib.it

Abstract. The ColorChecker dataset is the most widely used dataset for evaluating and benchmarking illuminant-estimation algorithms. Although it is distributed with a 3-fold cross-validation partitioning, no procedure is defined on how to use it. In order to permit a fair comparison between illuminant-estimation algorithms, in this short correspondence we define a fair comparison procedure, showing that illuminant-estimation errors of state-of-the-art algorithms have been underestimated by up to 33%. We also compute the lower error bounds that can be reached on this dataset, which demonstrates that the existing algorithms have not yet reached their maximum performance potential.

1 Introduction

The ColorChecker dataset (CC) is a benchmark dataset for illuminant estimation. It was first introduced by Gehler et al. in 2008 [1] and it is still widely used. It contains a total of 568 raw RGB images of indoor and outdoor scenes taken with two cameras, namely Canon 1D (86 images) and Canon 5D (482 images), and consists of typical photographic scenes. Over the years a linear re-processed version of the ColorChecker dataset has been released [2] and has become the reference version of the dataset.

The ground truth, i.e., the global illuminant color, associated with each image is defined as the RGB response from the achromatic patches of the eponymous Macbeth ColorChecker chart, that is placed in each scene and omitted during testing. The corners of the ColorChecker chart and those of all its patches were manually annotated. However, recent advancements in the literature (e.g., [3, 4]) have introduced automatic techniques that have demonstrated high accuracy in ground truth generation, suggesting that these methods may serve as viable alternatives to manual annotation in future studies on larger datasets and to ensure coherent annotation across datasets. Figure 1 shows some sample images from the ColorChecker dataset.

In order to properly compare the performance of different illuminant-estimation algorithms on a common ground, in 2018 Hemrit et al. [5, 6] joined efforts to establish a new, recommended ground truth (REC) and made it available to the research community.

Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Figure 1: Some sample images within the CC dataset (gamma corrected for better visualization)

Since its original version, the ColorChecker dataset has been distributed with a 3-fold cross-validation partitioning. Unfortunately, no procedure was given on how to use this partitioning, leading to a potential data leakage between training and test folds with researchers sometimes tweaking model hyper-parameters on the test set[1]. This might not have been a problem when illuminant estimation algorithms had only a few parameters/hyperparameters to tune (e.g., statistics-based methods [8] or parametric methods [9, 10]), but it has become a problem now that the vast majority of new algorithms are based on deep learning with a number of parameters to train ranging from a few thousands [11, 12, 13] up to several millions [7, 14, 15], thereby increasing the likelihood of overfitting to the test partition.

In this paper we define a fair comparison procedure for illuminant-estimation algorithms on the ColorChecker dataset. We show that the angular error statistics of state-of-the-art algorithms have been underestimated by up to about 33%. Under the assumption that the illuminant is spatially invariant, we also compute the lower bound errors that can be achieved on the ColorChecker dataset, showing that they have not been reached. Finally, we create a repository (`https://github.com/simone-255-255-255/fairCC`) reporting the performance of several state-of-the-art algorithms under the proposed fair-comparison procedure.

2 A fair comparison procedure

The suggested procedure for fair comparison of illuminant-estimation algorithms is reported in Figure 2. It begins with the original 3-fold cross-validation partitioning. In each cross validation round, two folds are used to train the model, while the third one is used exclusively for testing.

For those algorithms that need a validation set either to tune some hyperparameters, to select the best model, or to determine the termination epoch of the training, a validation set is created from the two training folds and excluding the test fold, thus preventing any data leakage and ensuring a more accurate estimate of the generalization error. To this end, for each cross-validation round, a further 80%-20% split is provided in the repository.

If the algorithm uses some forms of data augmentation in training, the unaugmented training set can be used for validation.

Since many illuminant estimation algorithms are stochastic, in order to avoid reporting cherry-picked results, we propose averaging the reported statistics over at least three independent runs. Moreover, as already commonly done, the ColorChecker chart in each image must be masked by setting the pixel values

[1] Official repository of [7], FAQ question (d): https://github.com/yuanming-hu/fc4. Last accessed May-2025.

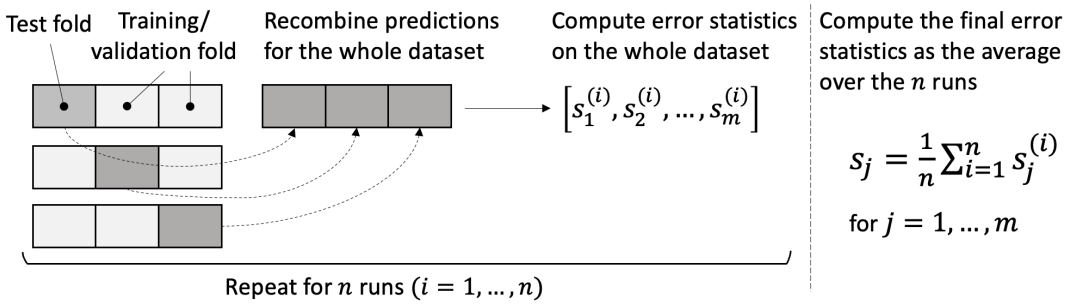

Figure 2: Schematic representation of the fair ColorChecker comparison procedure: it begins with the original 3-fold cross-validation partitioning. In each cross validation round, two folds are used to train the model, while the third one is used exclusively for testing. Predictions on the three test folds are recombined and error statistics are computed for the whole dataset. The training/testing procedure is repeated n times on the same split and the final error statistics are computed as the average over the n rounds to account for method's stochasticity.

Table 1: Performance comparison in terms of angular error on the ColorChecker dataset for several illuminant estimation algorithms with fair and unfair training or hyperparameters setting for some statistics-based algorithms (top) and some learning-based algorithms (bottom). Average and maximum percentage changes ($\Delta\%$) computed on all the unfair statistics with respect to corresponding fair ones are also reported.

	Fair								Unfair								$\Delta\%$ wrt Fair	
Method	Mean	Med.	Tri-m.	B-25	W-25	95-P	99-P	Max	Mean	Med.	Tri-m.	B-25	W-25	95-P	99-P	Max	Avg.	Max.
SoG [9]	4.07	2.70	3.14	0.55	9.79	12.20	17.00	21.89	3.95	2.54	3.01	0.56	9.62	12.44	16.53	20.48	-2.49%	-6.43%
GGW [9]	4.05	2.60	3.06	0.56	9.78	12.30	16.97	21.16	4.00	2.62	3.02	0.55	9.73	12.44	16.94	20.58	-0.57%	-2.74%
GE1 [9]	3.89	2.77	3.10	0.78	8.83	11.09	14.57	22.60	3.87	2.84	3.11	0.84	8.60	10.99	14.88	20.18	-1.21%	-10.68%
GE2 [9]	3.89	2.88	3.13	0.75	8.80	10.90	14.02	23.10	3.87	2.77	3.04	0.79	8.73	10.92	14.37	22.43	-1.25%	-4.23%
FFCC (model J) [16]	2.23	1.45	1.59	0.35	5.46	7.33	10.85	17.27	1.79	0.98	1.19	0.27	4.63	5.83	11.02	17.46	-16.49%	-32.52%
FC^4 [7] (avg 3 runs)	2.14	1.44	1.57	0.40	5.08	6.50	12.75	15.28	2.11	1.45	1.58	0.43	4.94	6.33	10.80	16.02	-1.02%	-15.29%
FC^4 [7] (best run)	2.05	1.33	1.46	0.38	4.95	6.19	11.98	15.56	1.99	1.31	1.47	0.42	4.74	6.48	9.76	14.31	-2.30%	-18.53%

to zero in RGB space before applying any illuminant-estimation algorithm, i.e., both in the training and in the testing phase.

3 Re-evaluation of illuminant estimation algorithms

Since in illuminant estimation it is more important to measure the color of the illuminant rather than its magnitude, illuminant estimation performance is usually measured in terms of angular error between the RGB values of the estimated illuminant i_{est} and the RGB values of the ground truth illuminant i_{gt} [17, 18]. The recovery angular error err_{rec} is therefore defined as:

$$err_{rec} = \cos^{-1}\left(\frac{i_{gt} \cdot i_{est}}{||i_{gt}||\ ||i_{est}||}\right) \tag{1}$$

Table 1 reports the performance in terms of recovery angular error on the ColorChecker dataset for several illuminant estimation algorithms when hyperparameters/parameters are trained in the proposed fair way, and they are compared to the performance when hyperparameters/parameters are trained in an unfair way. For the unfair results we intentionally cause a data leakage: for statistics-based and parametric methods we optimize/train the algorithms based on the test fold performance, while for training-based methods and methods based on deep learning we use the test fold for best model selection. For the stochastically-trained FC^4 [7] we evaluate both a 3-run average version, and a best-run version.

For both fair and unfair training setups we report several commonly used angular error statistics: average, median, tri-mean, best 25%, worst 25%, 95th percentile, 99th percentile, and maximum. In order to compare the fair and unfair performance, the percentage changes ($\Delta\%$) of all the unfair statistics with respect to corresponding fair ones are computed, and their average and maximum values across the statistics considered are reported.

Table 2: Lower bound angular errors on the ColorChecker dataset: intra-patch (a) and inter-patch (b). We report per-image statistics (rows of each table), i.e. how the distances between pseudo and REC ground truths are summarized per image, as well as dataset statistics (columns of each table).

	Dataset-level statistics							
Image-level statistics	Mean	Med.	Tri-m.	B-25	W-25	95-P	99-P	Max
Mean	0.057	0.045	0.048	0.023	0.109	0.129	0.209	0.377
Median	0.054	0.044	0.045	0.021	0.105	0.132	0.186	0.419
Max	0.095	0.075	0.079	0.037	0.189	0.224	0.391	0.572
Std	0.029	0.022	0.023	0.010	0.060	0.075	0.131	0.214

(a)

	Dataset-level statistics							
Image-level statistics	Mean	Med.	Tri-m.	B-25	W-25	95-P	99-P	Max
Mean	0.92	0.68	0.77	0.39	1.76	2.09	2.92	4.78
Median	0.61	0.46	0.50	0.30	1.19	1.52	2.51	4.59
Max	2.15	1.42	1.71	0.71	4.66	5.57	6.99	9.95
Std	0.78	0.48	0.60	0.21	1.80	2.20	2.83	4.05

(b)

From the results reported in Table 1 we can observe how an unfair use of the ColorChecker dataset can lead to underestimating the angular error statistics for simple statistics-based algorithms by up to 3% on average, while in the worst case, a single angular error statistic can be underestimated by about 11%. For learning-based algorithms, the error statistics on average can be underestimated by as much as 16%. In the worst case instead, a single angular error statistic can be underestimated up to 32%.

4 Computation of the lower bound errors on the ColorChecker dataset

In this section, we assess whether or not recent illuminant-estimation algorithms have reached the lower bound of the possible error on the ColorChecker dataset. This lower bound error is computed under the assumption that the illumination in each image is spatially invariant, although this assumption is violated to a certain extent in almost every image of any existing dataset categorized as single illuminant due to the presence of shadows and inter-reflections [19].

Two different analyses are carried out with the aim of measuring how noisy the ColorChecker ground truth is: an intra-patch and an inter-patch analysis.

4.1 Intra-patch analysis

In the intra-patch analysis, we divide the neutral patch selected as REC ground truth (i.e., the brightest neutral patch not containing any clipped pixel) for each image into five sub-patches having the same size: four non-overlapping sub-patches, and one central sub-patch having the same size. From each sub-patch a new illuminant is extracted. Therefore, for each image, we have the REC ground truth and five additional pseudo-ground-truth illuminants. The intra-patch analysis procedure is depicted in the first image in the top row of Figure 3. The angular errors between each pseudo-ground-truth illuminant and the REC ground truth are computed, and for each image several statistics are computed: average, median, maximum, and standard deviation. These statistics are accumulated over the whole ColorChecker dataset and summarized using the same error statistics used in the previous section. The results are reported in Table 2(a).

4.2 Inter-patch analysis

In the inter-patch analysis, we consider all the neutral patches that are not clipped according to [2] (i.e., they have no digital count higher than 3300). Then, from each not-clipped patch, a new illuminant is extracted. Therefore, for each image we have the REC ground truth and up to five pseudo-ground-truth illuminants. The inter-patch analysis procedure is depicted in the first image in the bottom row of Figure 3. The same procedure carried out for the intra-patch case is then applied: the angular errors between

each pseudo-ground-truth illuminant and the REC ground truth are computed, image-level statistics are accumulated, summarized at dataset level, and reported in Table 2(b). The results show that only the maximum distance in the inter-patch setup has error statistics close to those of recent illuminant-estimation algorithms.

4.3 Actual distance from lower bound errors

In order to measure how far the current results are from the lower bound errors, for each image we project the intra-patch and inter-patch pseudo-ground-truths on the Maxwell triangle with chromaticities:

$$\begin{cases} x = \sqrt{3}(r - g) \\ y = r + g - 2b \end{cases} \tag{2}$$

together with the REC ground truth, and compute their respective convex hulls. A visual summary of the complete procedure is shown in Figure 3 for both the intra-patch and inter-patch analyses. We then compute the ratio of images in the ColorChecker dataset for which the illuminant estimated by an algorithm falls within these convex hulls. The results are reported in Table 3.

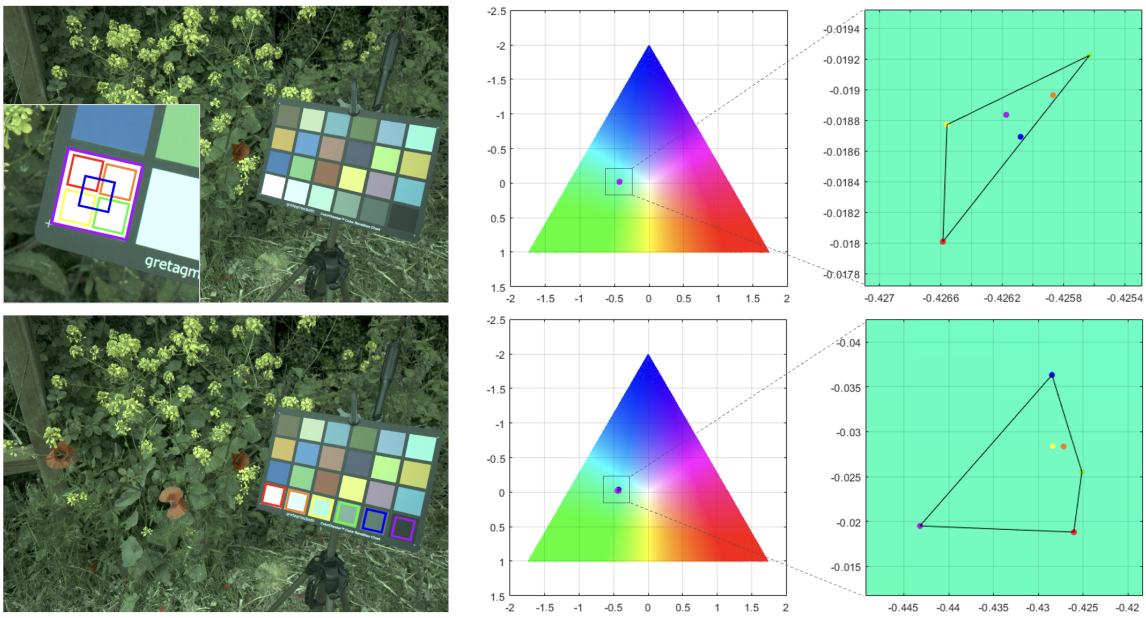

Figure 3: Example of convex hull computation in the Maxwell triangle for REC and pseudo ground truths: intra-patch (top), inter-patch (bottom) for an image in the ColorChecker dataset. For each row the first image shows six regions used to compute the different ground truths. The second image maps the computed ground truths in the chromaticity space defined by Maxwell triangle, represented as color dots with the same color coding used to highlight the ground truth regions. The third image is a zoomed version of the second one, to better show the ground truth chromaticities, with a black line showing their computed convex hull.

We observe that for the intra-patch configuration, the lower bound errors are reached in at most 0.53% of the images even considering the unfair setup. In the inter-patch configuration instead, the lower bound errors are reached in at most 7.22% of the images.

The results thus confirm that current illuminant-estimation algorithms are still far from reaching the lower bound errors on the ColorChecker dataset, and thus ColorChecker is still distant from its end of life.

Further analysis could also be performed by counting in how many images the estimated illuminant is equally physically possible as the ground truth, i.e., how many times it belongs to the feasible set of illuminants [20].

Table 3: Ratio of images of the ColorChecker dataset on which the estimated illuminant falls within the convex hull of the pseudo and REC ground truths.

Method	Intra-patch Fair	Intra-patch Unfair	Inter-patch Fair	Inter-patch Unfair
SoG [9]	0.18%	0.18%	2.11%	2.29%
GGW [9]	0.00%	0.00%	2.64%	2.64%
GE1 [9]	0.00%	0.00%	0.53%	0.70%
GE2 [9]	0.35%	0.00%	0.53%	0.35%
FFCC (model J) [16]	0.35%	0.53%	5.46%	4.40%
FC^4 [7] (average over 3 runs)	0.06%	0.23%	4.28%	5.63%
FC^4 [7] (best run)	0.00%	0.00%	4.75%	7.22%

5 Conclusion

The ColorChecker dataset is a benchmark dataset for illuminant-estimation algorithms that is distributed with a 3-fold cross-validation partitioning. Unfortunately, no standardized procedure exists on how to use it, making the comparison of different algorithms problematic.

In order to permit a fair comparison, in this paper we define a fair comparison procedure for illuminant-estimation algorithms on the ColorChecker dataset. We re-evaluate the performance of several state-of-the-art algorithms, showing that the angular error statistics were underestimated by up to about 33%.

We created a public repository to report the fair performance of more algorithms in the state of the art.

Finally, we also computed the lower bound errors that can be reached on the ColorChecker dataset by measuring how noisy is the ColorChecker ground truth, under the assumption of spatially invariant scene illumination. From our analysis we observe that the lower bound errors are reached in at most 0.35% and 5.46% of the images in the intra-patch and inter-patch configuration respectively.

References

[1] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[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: May-2025).

[3] K. Hirakawa, "Colorchecker finder," https://sites.google.com/a/udayton.edu/issl/software/macbeth-colorchecker-finder, (accessed: May-2025).

[4] L. Cogo, M. Buzzelli, S. Bianco, and R. Schettini, "Robust camera-independent color chart localization using yolo," *Pattern Recognition Letters*, vol. 192, pp. 51–58, 2025.

[5] G. Hemrit, G. D. Finlayson, A. Gijsenij, P. Gehler, S. Bianco, B. Funt, M. Drew, and L. Shi, "Rehabilitating the colorchecker dataset for illuminant estimation," in *IS & T/SID Color Imaging Conference*, 2018, p. 350 – 353.

[6] G. Hemrit, G. D. Finlayson, A. Gijsenij, P. Gehler, S. Bianco, M. S. Drew, B. Funt, and L. Shi, "Providing a single ground-truth for illuminant estimation for the colorchecker dataset," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 5, pp. 1286–1287, 2019.

[7] Y. Hu, B. Wang, and S. Lin, "Fc4: Fully convolutional color constancy with confidence-weighted pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4085–4094.

[8] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Josa*, vol. 61, no. 1, pp. 1–11, 1971.

[9] J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Transactions on image processing*, vol. 16, no. 9, pp. 2207–2214, 2007.

[10] D. Cheng, D. K. Prasad, and M. S. Brown, "Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution," *JOSA A*, vol. 31, no. 5, pp. 1049–1058, 2014.

[11] H. Gong, "Convolutional mean: A simple convolutional neural network for illuminant estimation," *30th British Machine Vision Conference 2019, BMVC 2019*, 2020.

[12] I. Domislović, D. Vršnak, M. Subašić, and S. Lončarić, "One-net: Convolutional color constancy simplified," *Pattern Recognition Letters*, vol. 159, pp. 31–37, 2022.

[13] M. Buzzelli and S. Bianco, "A convolutional framework for color constancy," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[14] S. Bianco and C. Cusano, "Quasi-unsupervised color constancy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 212–12 221.

[15] H. Yu, K. Chen, K. Wang, Y. Qian, Z. Zhang, and K. Jia, "Cascading convolutional color constancy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 725–12 732.

[16] J. T. Barron and Y.-T. Tsai, "Fast fourier color constancy," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 886–894.

[17] S. D. Hordley and G. D. Finlayson, "Reevaluation of color constancy algorithm performance," *JOSA A*, vol. 23, no. 5, pp. 1008–1020, 2006.

[18] G. D. Finlayson, R. Zakizadeh, and A. Gijsenij, "The reproduction angular error for evaluating the performance of illuminant estimation algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1482–1488, 2016.

[19] L. Xu and B. Funt, "How multi-illuminant scenes affect automatic colour balancing," in *Proc. AIC 2015 International Colour Association Conference*. Simon Fraser University, 2015.

[20] D. A. Forsyth, "A novel algorithm for color constancy," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 5–35, 1990.