A Multi-Hypothesis Approach to Color Constancy: supplementary material

Daniel Hernandez-Juarez[1], Sarah Parisot[1,2], Benjamin Busam[1,3], Aleš Leonardis[1]
Gregory Slabaugh[1], Steven McDonagh[1]

dhernandez0@gmail.com,
{sarah.parisot, benjamin.busam, ales.leonardis, gregory.slabaugh, steven.mcdonagh} @huawei.com

[1]Huawei Noah's Ark Lab [2]Mila Montréal [3]Technical University of Munich

We provide additional material to supplement our main paper. In Section 1, we present our shallow CNN architecture. Two experimental studies on the number of illuminant candidates are provided in Section 2. In Section 3, we report details on NUS [4] per-camera median angular error to provide evidence for our claim that we consistently improve accuracy for each camera, using multi-camera training (see main paper Section 4.4). In Section 4, we show additional results from our exploration of candidate selection strategy. Section 5 provides run-time measurements and in Section 6 we observe failure cases and discuss limitations of our method. Finally, Section 7 provides additional visual results comparing our method with FFCC [3].

1. Architecture details

In Table 1, we present our CNN architecture. We propose a shallow CNN, one spatial 3×3 convolution and two subsequent layers constituting 1×1 convolutions with a final global spatial pooling. Lastly, three fully connected layers gradually reduce the dimensionality to one.

Layer	Kernel	Input	Output
Conv.	3×3	64×64×3	64×64×64
Conv.	1×1	64×64×64	64×64×64
Conv.	1×1	64×64×64	64×64×128
Avg. Pool.	64×64	64×64×128	128
FC	-	128	64
FC	-	64	32
FC	-	32	1

Table 1. CNN architecture details. Fully connected layers and convolutions are followed by a ReLU activation except the last layer.

2. Number of illuminant candidates

In Table 2 we present a study varying the number of candidate illuminants produced by K-means. We find experimentally that accuracy improves with the number of cluster centres until a plateau is reached, suggesting that we need $\sim$100 candidate illuminants to achieve competitive angular error for the Gehler-Shi dataset [12, 6].

# candidates	Mean	Med.	Tri.	Best 25%	Worst 25%
5	2.79	2.06	2.20	0.67	6.23
25	2.24	1.50	1.64	0.38	7.34
50	2.25	1.47	1.66	0.37	5.51
100	2.15	1.38	1.55	0.40	5.16
120	2.10	1.32	1.53	0.36	5.10
150	2.16	1.33	1.53	0.39	5.25
200	2.16	1.39	1.59	0.37	5.20

Table 2. Error for differing number of candidates for K-means candidate selection. Angular error for Gehler-Shi dataset [12, 6].

# candidates	Mean	Med.	Tri.	Best 25%	Worst 25%
5	2.53	1.71	1.81	0.51	6.06
25	2.28	1.43	1.59	0.45	5.63
50	2.28	1.46	1.61	0.46	5.52
100	2.12	1.31	1.45	0.40	5.31
120	2.07	1.31	1.43	0.41	5.12
150	2.16	1.32	1.49	0.40	5.34
200	2.12	1.33	1.47	0.40	5.27

Table 3. Angular error for the Cube challenge [9] trained only on NUS [4] and Gehler-Shi [12, 6]. For our method, candidate selection is performed on Cube+ [1] with varying K for K-means candidate selection.

Additionally, we provide analogous results for different values of K for K-means candidate selection for the training-free model (see main paper Section 4.5), in Table 3. We observe stability for $K >= 25$. The low number of candidates required is likely linked to the two Cube datasets having reasonably compact illuminant distributions.

3. NUS per-camera median angular error

We provide evidence supporting our paper claim that training the proposed model with images from multiple cameras outperforms individual, per-camera, model training (see Section 4.4, of the main paper).

We reiterate that folds are divided such that scene content

Camera	Ours (one model per device)	Ours (multi-device training)
Canon EOS-1Ds Mark III	1.59	1.49
Canon EOS 600D	1.49	1.23
Fujifilm X-M1	1.34	1.33
Nikon D5200	1.69	1.50
Olympus E-PL6	1.30	1.13
Panasonic Lumix DMC-GX1	1.43	1.21
Samgsung NX2000	1.54	1.42
Sony SLT-A57	1.50	1.41

Table 4. Median angular error of our method for each individual camera of NUS [4].

is consistent within a fold, across all cameras. This ensures to avoid testing on familiar scene content, as observed by a different camera during training. Towards reproducibility, and fair comparison, our suppplementary material provides the cross validation (CV) splits, used in the main paper, for multi-device training. CV splits were generated manually by ensuring that all images of the same scene (across different cameras) belong to the same fold.

In Table 4 we report median angular-error for test images of the NUS [4] dataset. Multi-device training can be seen to consistently improve the median angular error for all NUS cameras at test time.

4. Candidate selection methods

We report additional illuminant candidate selection strategies explored during our investigation.

Uniform-sampling: we consider the global extrema of our measured illuminant samples (max. and min. in each color space dimension) and sample n points uniformly using an $[\frac{r}{g}, \frac{b}{g}]$ color space. These samples constitute our illuminant candidates.

K-means clustering: cluster centroids define candidates, as detailed in the main paper, Section 3.2 and other recent color constancy work [10]. We use RGB color space for clustering, and experimentally verified that both $[\frac{r}{g}, \frac{b}{g}]$ and RGB color spaces provided similar accuracy.

Mixture Model (GMM): we fit a GMM to our measured illuminant samples in $[\frac{r}{g}, \frac{b}{g}]$ color space, and then draw n samples from the GMM to define illuminant candidates.

We use 121 candidates (11×11 grid) for uniform candidate selection. For GMM candidate selection, we fit 10 two-dimensional Gaussian distributions and sample 120 candidates.

In Table 5 we report inference performance on the Cube challenge [9] data set using the described candidate selection strategies. We observe that simple uniform-sampling candidate selection performs reasonably well. The strategy provides an extremely simple implementation yet, by definition, will also sample some portion of very unlikely candidates. We note, however, that if the interpolation between candidates span the illuminant space, our method can learn to interpolate these candidates appropriately, accounting for this. The GMM approach also results in slightly weaker accuracy performance *c.f.* K-means, motivating our choice of sampling strategy in the experimental work for the main paper.

Method	Mean	Med.	Tri.	Best 25%	Worst 25%
Uniform	2.11	1.20	1.30	0.41	5.45
GMM	2.27	1.10	1.25	0.41	6.31
K-means	1.99	1.06	1.14	0.35	5.35

Table 5. Angular error for Cube challenge [9] of our method using different candidate selection methods.

5. Inference run-time

We report inference run-time results for the Gehler-Shi dataset [12, 6] in Table 6. We note that our real-time inference speed is obtained using a Nvidia *Tesla V100* card and unoptimised implementation (PyTorch 1.0 [11]). We highlight that our algorithm is highly parallelizable, each illuminant candidate likelihood can be computed independently, however, we obtain the run-time with single-thread implementation. Our input image resolution is 64×64 and timing results are recorded using K-means candidate selection with K=120. The timing performance of other methods are obtained from their respective citations. We acknowledge that timing comparisons are non-rigorous; reported run-times are measured using differing hardware. To provide additional fair comparison; Table 7 reports run-times for both our method and the official[1] FFCC [3] implementation run on Matlab R2019b, under common hardware (Intel Core i9-9900X (3.50GHz)).

Method	Run-time (ms)	Hardware
CCC [2]	520	2012 HP Z420 workstation (CPU)
Cheng *et al.* 2015 [5]	250	Intel Xeon 3.5GHz (CPU)
FC4 [8]	25	Nvidia GTX TITAN X Maxwell (GPU)
FFCC [3] (model Q)	1.1	Intel Xeon CPU E5-2680 (CPU)
CM 2019 [7]	1	Nvidia Tesla K40m (GPU)
Ours	7.3	Nvidia Tesla V100 (GPU)

Table 6. Inference time for images of Gehler-Shi dataset [12, 6]. Run-time is provided in milliseconds (ms).

Method	Run-time (ms)
FFCC [3] (model Q)	1.2
Ours	128

Table 7. Inference time for images of Gehler-Shi dataset [12, 6]. Run-time is provided in milliseconds (ms). Run-time measured using a Intel Core i9-9900X (3.50GHz) CPU.

6. Failure cases

In Figures 1 to 3 we provide observed limitations and failure cases. Our method learns to interpolate between candidate illuminants, that are observed during training, but not

[1] https://github.com/google/ffcc

(a) Our prediction (angular-error = $6.14°$)

(b) Ground Truth

Figure 1. This scene can be observed to be illuminated by more than one light source, breaking the single global illuminant assumption. Images are rendered in sRGB color space.

(a) Our prediction (angular-error = $6.05°$)

(b) Ground Truth

Figure 2. An ambiguous scene with multiple *plausible* solutions, highlighting the ill-posed nature of the color constancy problem. Our method infers a plausible, yet incorrect, solution; that the color of the stone building is white. Images are rendered in sRGB color space.

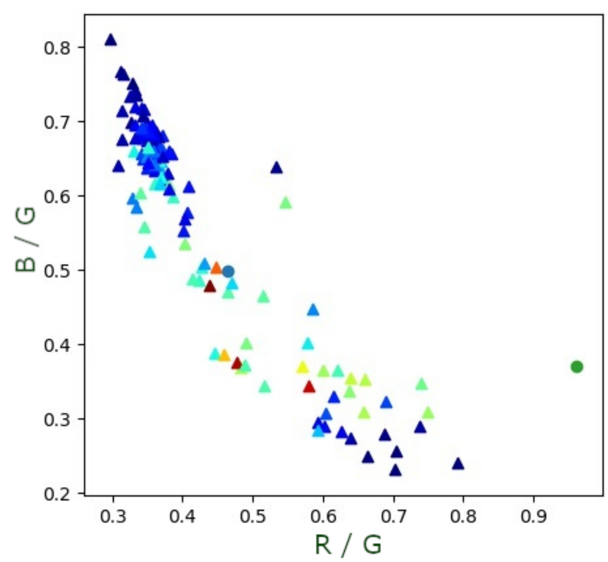
to extrapolate to new illuminants. In Figure 3c, the ground truth illuminant (green filled circle) is clearly out of distribution, with no similar candidate illuminants observed during training. The resulting inference accuracy in Figure 3a suffers as a result.

Further, our single global illuminant assumption can be seen to be violated in Figure 1. The predicted illuminant attempts to balance the outer boundary portions of the wall painting as achromatic, clearly illuminated from above (out of shot). The measured ground truth illuminant captures the desk lamp illumination, resulting in high angular error for this image due to the global assumption.

Finally, in Figure 2, we observe an example scene with extreme ambiguities. Our method appears to infer that the stone building in the scene background is achromatic, producing a highly plausible image. Yet the measured ground-truth illuminant illustrates the true building color to be of mild beige-yellow.

(a) Our prediction (angular-error = $20.12°$)

(b) Ground Truth



(c) $\frac{r}{g}$, $\frac{b}{g}$ plot of candidates

Figure 3. This challenging scene is illuminated by a measured illumination color not seen during training. In Figure 3c the green circular point corresponds to the ground-truth illuminant and can be observed to be outwith the illuminant candidate distribution. Images are rendered in sRGB color space.

7. Additional qualitative results

In Figure 4, we provide additional qualitative results in the form of test images from the NUS [4] dataset (Sony camera). For each test sample we show the input image and a white-balanced image, corrected using the ground-truth illumination in addition to the output of our model ("multi-device training + pretraining"), and that of FFCC (model Q) [3]. Each row consists of: (a) the input image (b) FFCC [3] (c) our prediction (d) ground truth.

In similar fashion to [2], we adopt the strategy of sorting test images by the combined mean angular-error of the two evaluated methods. We present images of increasing average difficulty, sampled with a uniform spacing. Images are corrected by inferred illuminants, applying an estimated CCM (Color Correction Matrix), and standard sRGB gamma correction. The *Macbeth Color Checker* is used to generate the ground-truth and is present in the images, however the relevant regions are masked during both training and inference. It can be observed in Figure 4 in almost all sampled cases, we see consistently improved results with our approach.

We provide further extremely challenging examples in Figure 5. We explicitly select the five largest combined mean angular-error images. We observe that our method shows consistently strong performance and also highlight that these samples constitute cases of both ambiguous and multi-illuminant scenes, breaking the fundamental global illuminant assumption (made by both methods).

References

[1] Nikola Banic and Sven Loncaric. Unsupervised learning for color constancy. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 4: VISAPP, Funchal, Madeira, Portugal, January 27-29, 2018*, pages 181–188, 2018.

[2] Jonathan T. Barron. Convolutional color constancy. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 379–387, 2015.

[3] Jonathan T. Barron and Yun-Ta Tsai. Fast fourier color constancy. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6950–6958, 2017.

[4] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*, 31(5):1049–1058, 2014.

[5] Dongliang Cheng, Brian L. Price, Scott Cohen, and Michael S. Brown. Effective learning-based illuminant estimation using simple features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1000–1008, 2015.

[6] Peter V. Gehler, Carsten Rother, Andrew Blake, Thomas P. Minka, and Toby Sharp. Bayesian color constancy revisited. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[7] Han Gong. Convolutional mean: A simple convolutional neural network for illuminant estimation. In *Proceedings of the British Machine Vision Conference 2019, BMVC 2019, Cardiff University, Cardiff, UK, September 9-12, 2019*, 2019.

[8] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc^4: Fully convolutional color constancy with confidence-weighted pooling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 330–339, 2017.

[9] Karlo Koscevic and Nikola Banic. ISPA 2019 Illumination Estimation Challenge. `https://www.isispa.org/illumination-estimation-challenge`. Accessed November 14, 2019.

[10] Seoung Wug Oh and Seon Joo Kim. Approaching the computational color constancy as a classification problem through deep learning. *Pattern Recognition*, 61:405–416, 2017.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019.

[12] Lilong Shi and Brian Funt. Re-processed version of the gehler color constancy dataset. `https://www2.cs.sfu.ca/~colour/data/shi_gehler/`. Accessed November 14, 2019.

(a) Input image (b) FFCC (error: 0.08°) (c) Ours (error: 0.07°) (d) Ground Truth

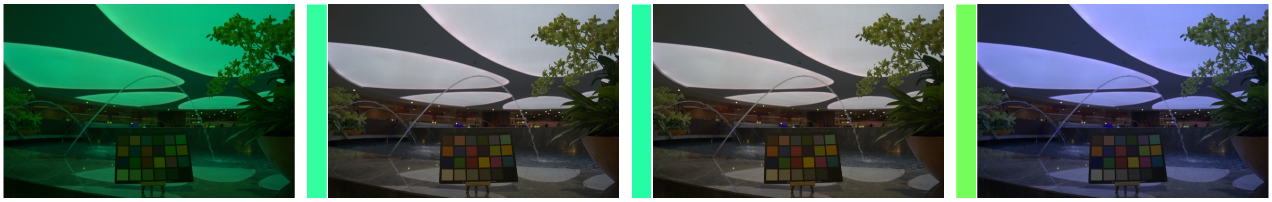
(a) Input image (b) FFCC (error: 1.66°) (c) Ours (error: 0.18°) (d) Ground Truth

(a) Input image (b) FFCC (error: 1.66°) (c) Ours (error: 1.41°) (d) Ground Truth

(a) Input image (b) FFCC (error: 2.68°) (c) Ours (error: 2.50°) (d) Ground Truth

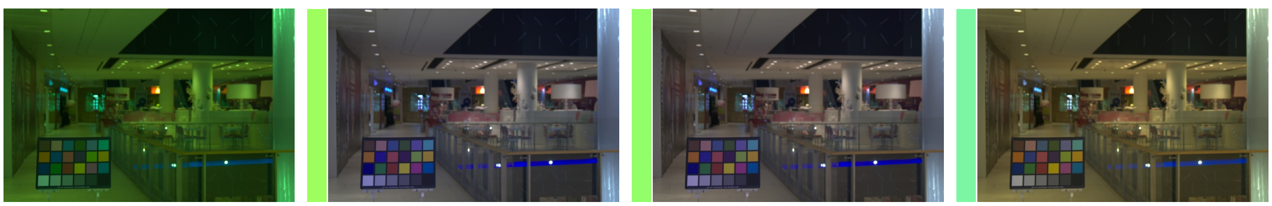
(a) Input image (b) FFCC (error: 11.93°) (c) Ours (error: 20.12°) (d) Ground Truth

Figure 4. Visual comparisons of FFCC [3] and our method. We sort test results of the Sony dataset (NUS [4]) by the combined (sum total) mean angular error of the two evaluated methods and then uniformly sample images to select test images. Images are rendered in sRGB color space.



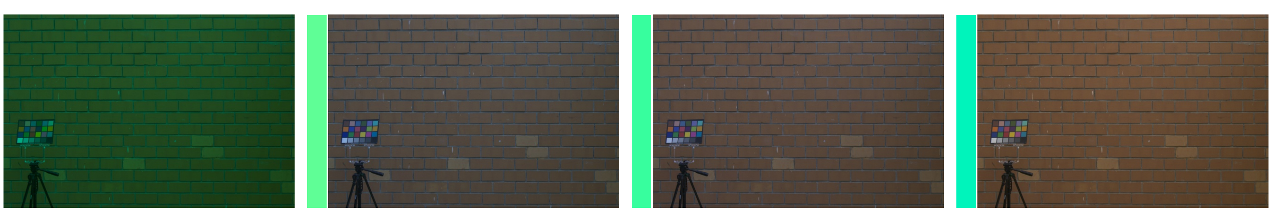
(a) Input image (b) FFCC (error: 11.87°) (c) Ours (error: 13.17°) (d) Ground Truth

(a) Input image (b) FFCC (error: 16.66°) (c) Ours (error: 7.24°) (d) Ground Truth



(a) Input image (b) FFCC (error: 12.27°) (c) Ours (error: 10.23°) (d) Ground Truth

(a) Input image (b) FFCC (error: 9.36°) (c) Ours (error: 11.21°) (d) Ground Truth



(a) Input image (b) FFCC (error: 10.37°) (c) Ours (error: 7.49°) (d) Ground Truth

Figure 5. Visual comparison of FFCC [3] and our method with Sony dataset (NUS [4]). We select the five **largest** combined mean angular error to explore method behaviour for images that are commonly challenging. Images are rendered in sRGB color space.