https://doi.org/10.2352/issn.2169-2629.2019.27.21

Evaluating Colour Constancy on the new MIST dataset of Multi-Illuminant Scenes

Xiangpeng Hao, Brian Funt, Hanxiao Jiang;
School of Computing Science, Simon Fraser University; Vancouver, Canada

Abstract

A new image test set of synthetically generated, full-spectrum images with pixelwise ground truth has been developed to aid in the evaluation of illumination estimation methods for colour constancy. The performance of 9 illumination methods is reported for this dataset along and compared to the optimal single-illuminant estimate. None of the methods specifically designed to handle multi-illuminant scenes is found to perform any better than the optimal single-illuminant case based on completely uniform illumination.

Introduction

Until relatively recently, the majority of work on colour constancy has been based on the assumption that the scene is lit by a single illuminant. Everyone is aware that this assumption is very often violated, but there are only very limited image datasets, such as those created by Beigpour et al. [1] [2], with ground truth data for multi-illuminant scenes. However, these data sets are both small and mainly consist of images of quite constrained single-object scenes. We introduce a new dataset of 1,000 full-spectrum images of complex scenes and test the publicly available implementations of colour constancy methods on them, including both those specifically designed for multi-illuminant scenes.

The goal of colour constancy is often defined as the recovery of the 'true' or 'intrinsic' colour of each surface in an image. However, as Logvinenko et al [3] point out, the problems of metamer mismatching make this formulation fundamentally incorrect. Surface reflectance is an inherent, intrinsic property of a surface, but colour—despite its use to that effect in everyday language—is not an intrinsic property of a surface. We can avoid this issue, however, since most colour constancy methods are based on a two-step process: (1) estimate the chromaticity of the illumination; (2) adjust the image colours relative to the chosen canonical 'white' illuminant using von Kries scaling. Recovering the chromaticity of the illumination is a well-posed problem, even though recovering the intrinsic surface colour is not. Fortunately, the difference between the majority of colour constancy methods lies in the illumination estimation (IE) step and so performance measures are based on evaluating the accuracy of the estimated chromaticity of the illuminant. Our focus here is on evaluating the IE performance of existing methods on multi-illuminant scenes having spatially-varying chromaticity of the lighting due both to the mixture of the light from various light sources, plus the lighting effects created by interreflections.

This paper's contributions are as follows.

1. We compare and analyse some representative IE methods on MIST. This is the first time these methods are evaluated with accurate, pixel-wise ground truth.

2. We provide detailed, in-depth evaluation results, including the output of each algorithm, a pixel-wise angular/distance error heatmap for each algorithm, and the optimal single-illuminant estimate.

3. We make a benchmarking system in Python that provides a uniform way to evaluate different IE methods. This system is designed to support any dataset in any image format, and similarly any algorithm implemented in any programming language. This generality is accomplished by allowing the user to write brief segments of wrapper code in Python that are then called by the benchmarking system itself. It is fully open-sourced and free for research use. It is available for download from: https://github.com/XiangpengHao/ColorConstancy.

Background

Beigpour et al. [2] provide a dataset of 600 multi-illuminant images photographed in a laboratory setting with ground truth data in terms of the surface colours under a specific white light. They built the dataset by very careful construction of each scene and camera setup. After capturing images of a scene under 20 different lighting conditions, the scene is spray painted grey and photographed again under the same 20 lighting conditions. This provides the necessary information for computing the ground truth. The scenes are quite limited in their complexity, usually of a single object. The final 600 images in the dataset are of 5 scenes under 20 different illumination conditions, photographed using 6 cameras. In other words, for a given camera there are only 100 images of 5 scenes/objects. Figure 1 shows two examples from that dataset.

Beigpour et al. [1] also generated a set of synthetic images of multi-illuminant scenes. The rendering was done with Photon mapping using 6 bands. The dataset is quite small, containing only 36 low resolution (480x270) images of 9 scenes under 4 different illumination conditions, 2 of which are more or less uniform in spectral power distribution. Hence, only 18 of these are truly multi-illuminant. Figure 1 (bottom) shows an example of a scene under two different illumination conditions along with the ground truth (middle).

Figure 1. *Example images from the real image (top) and synthetic image (bottom) data sets of Beigpour et al.* [2] [1]

The New MIST Multi-Illuminant Dataset

MIST [4] is a dataset containing 1,000 full-spectrum images created via ray-traced computer graphics rendering using a modified version of Blender [5]. The image data is stored in OpenEXR format using 16-bit floats. Figure 2 shows an example of an image from MIST. MIST is fully described elsewhere [4] and is available for download from: https://www2.cs.sfu.ca/~colour/data/

In summary, the image database has the following features.

(1) Full Spectrum Image Data

Every image is MIST is represented by full-spectrum pixel data ranging from 400nm to 695nm, sampled every 5nm. It also includes per-pixel depth information, although we do not make use of depth information here. For the evaluation of IE methods that work on standard linear sRGB data, we assume a colorimetrically accurate sRGB ideal camera (no noise, lens flare or blurring) model. In particular, we convert each spectrum to linear sRGB using the CIE 2° standard colorimetric observer colour matching functions to obtain the corresponding XYZ under CIE D65, and then convert XYZ to linear sRGB using the transformation matrix (Eq. 1) from the sRGB standard [6].

$$\begin{bmatrix} R_{linear} \\ G_{linear} \\ B_{linear} \end{bmatrix} = \begin{bmatrix} 3.2406 & -1.5372 & -0.4986 \\ -0.9689 & 1.8758 & 0.0415 \\ 0.0557 & -0.2040 & 1.0570 \end{bmatrix} \begin{bmatrix} X_{D65} \\ Y_{D65} \\ Z_{D65} \end{bmatrix} \quad (1)$$

(2) Pixel-By-Pixel Ground Truth

Most existing datasets provide only a single illuminant chromaticity as the ground truth value under the assumption that the scene is lit by only one illuminant. However, Xu et al. [7] found that almost half the images in the widely-used Gehler-Shi [8] [9] set, were, in fact, of multi-illuminant scenes. In other words, a single ground truth value must be incorrect for many locations in each such image. By using photo-realistic, multi-bounce ray-traced rendering, MIST is able to provide pixel-by-pixel ground truth data regarding both the percent spectral reflectance and the spectrum of the incident light. In terms of a ground truth represented in terms of sRGB, a canonical 'white' illuminant is chosen—usually either the equal-energy illuminant or CIE D65—to 'light' the spectral reflectances. Similarly, for the incident light the illuminant chromaticity is computed based on the light being reflected from an ideal reflector. Using this dataset, we can evaluate IE algorithms on scenes that include multiple illuminants, interreflections and transparency.

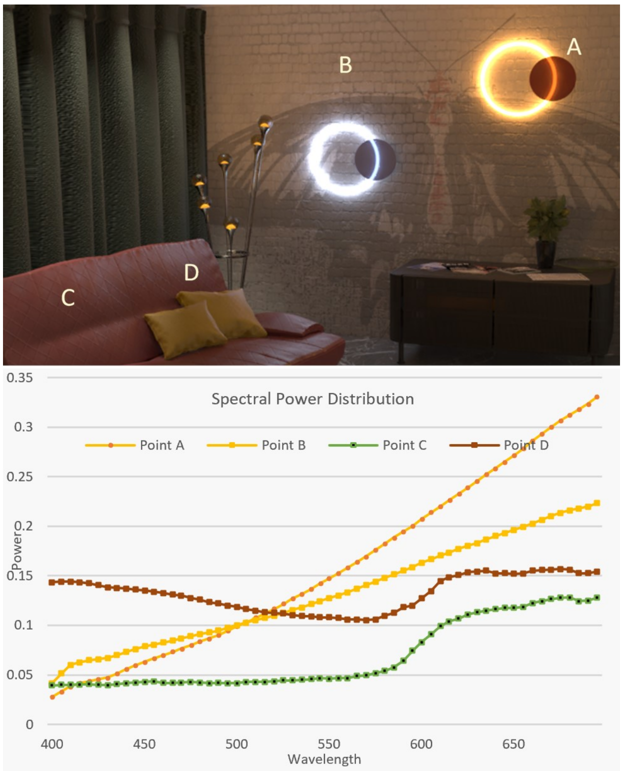

***Figure 2.** (top) Example of a full-spectrum image in which each pixel is represented by the spectral power distribution of the light reflected from the corresponding surface location. (bottom) Plots the spectral power distributions of the reflected light from the four points indicated in the image.*

(3) Physically Accurate Color Computation

MIST is a ray-tracing-based, synthetic dataset. Images are rendered one wavelength sample at a time in order to provide a physically accurate result. Standard RGB renderers model reflectances and illuminants as 3-tuples. For reflectances,

$$\rho_k^S = \int_w S(\lambda) R_k(\lambda) dx \quad (2)$$

For illuminants,

$$\rho_k^E = \int_w E(\lambda)\, R_k(\lambda)\, dx \quad (3)$$

Light reflected off a matte surface is then approximated as:

$$\rho_k^{E,S} = \rho_k^E \rho_k^S \quad (4)$$

However, Eq. (4) only approximates the physics of matte reflection. The correct model is:

$$\rho_k^{E,S} = \int_w E(\lambda) S(\lambda) R_k(\lambda) dx \quad (5)$$

Eq. (4) usually produces pleasing results, but not ones that are physically correct. As a result, standard photo-realistic renderings are not sufficient for evaluating IE methods. By rendering an 'image' of each wavelength separately, we can then compute the resulting color using Eq. 5 (discretized), rather than Eq. 4.

In Blender, specular reflection (gloss) is modeled with an additive component of the same spectral power distribution as of the incident illuminant. In other words, the interface reflection is the same for all wavelengths.

Illumination Estimation Error Analysis

We report IE errors based on the following measures.

rgb-Angular Error [10] per pixel 3-tuple ρ

$$err_{recovery} = \cos^{-1}\left(\frac{(\rho^E \cdot \rho^{Est})}{|\underline{\rho}^E|\,|\underline{\rho}^{Est}|}\right) \tag{6}$$

Reproduction Angular Error [10] per pixel

$$err_{reproduction} = \cos^{-1}\left(\underline{w}^{Est} \cdot \underline{w}\right) \tag{7}$$

where $\underline{w}^{Est} = \frac{\rho^E/\rho^{Est}}{|\rho^E/\rho^{Est}|}$ and $\underline{w} = \frac{\rho^E/\rho^E}{\sqrt{3}}$

rg-Distance Error per pixel

$$e = \{(r_E - r_{Est})^2 + (g_E - g_{Est})^2\}^{1/2} \tag{8}$$

where $r = R/(R+G+B)$ and $g = G/(R+G+B)$.

The mean and medians of these error measures taken over all image pixels are also reported.

Illumination Estimation Methods

IE methods are often divided into two categories; namely, static versus learning, as suggested by Gijsenij et al. [11]. The first type can be applied to any image without prior knowledge of the image's origin. In other words, no training is required. The second type of algorithm requires that the model be trained on a dataset of images from the same camera or cameras as the test image.

In our experiments, the following IE methods are evaluated over the whole MIST dataset. Many other methods have been described in the literature, but these are the ones that are either simple to implement or have implementations publicly available for download that actually worked when we tried to run them.

Max-RGB estimates the illuminant chromaticity as [max(R), max(G)]/[max(R)+max(G)+max(B)], where the maximum is across all pixels.

Greyworld estimates the illuminant chromaticity as [mean(R), mean(G)]/[mean(R)+mean(G)+mean(B)], where the mean is across all pixels.

Shades of Grey [12] method combines the above two methods by replacing the norm with the Minkowski p-norm and estimates the illuminant chromaticity as [pnorm(R), pnorm(G)] / [pnorm(R)+pnorm(G)+pnorm(B)].

Mixed K-Means [13] aims to solve the IE problem for multi-illuminant environments. It applies K-means to segment the image based on there being differing illuminants and then uses Max-RGB for IE within each segment.

Conditional GANs [14] [15] applies the pix2pix algorithm [16] to estimate the multiple illuminants in a non-uniform illumination environment.

Color by Colorization [17] is based on deep convolutional neural networks and applies a modification of the standard colorization of grayscale images to colored ones as a way to infer what the surface colors are likely to be under the canonical illuminant without explicitly estimating the illumination incident at each pixel.

FC4 [18] This method uses a fully convolutional network architecture in which patches throughout an image are assigned different confidence weights according to the value they provide for IE estimation. These weights are learned and applied within a pooling layer to merge them into a single global estimate.

Grey Edge [19] assumes that the average of the reflectance differences across a scene is achromatic.

Weighted Grey Edge [20] exploits distinct edge types to improve the performance of the original Grey Edge method. A variable weighting schema alternatingly estimates the chromaticity of the light source and updates the computed edge weights.

Experimental Setup and Evaluation

Preprocessing to PNG format

While most IE methods can be applied directly to OpenEXR 16-bit images, some learning-based methods only accept images in the same format as they are trained. For these methods, we converted the original OpenEXR MIST images to 8-bit PNG images. However, the mapping from OpenEXR to PNG is non-trivial because of the resulting reduction in dynamic range. Since this mapping is not the focus of this paper, we choose simply to perform a linear mapping between images in these two formats. However, to avoid the resulting images from being too dark, up to 5% of the pixels are clipped to 255. For image display, such a mapping would usually include a non-linear gamma correction component, but generally IE methods are tested on linear image data so a linear mapping between the two formats is required.

Oracle Method: Best Single-Illuminant Estimate

To provide a baseline with which to compare the results of IE methods applied to multi-illuminant images, we introduce the "Oracle" method. Given the corresponding ground truth image, Oracle finds the best single-illuminant chromaticity minimizing the rg-distance error in chromaticities averaged over the entire multi-illuminant input image. In other words, the Oracle provides the best possible single-illuminant estimate. Calculating the Oracle estimate is formulated using the following error metric:

$$Err = \left\|\vec{\gamma}\vec{\lambda} - \vec{\tau}\right\| \tag{9}$$

where $\vec{\gamma}$ is the vectorized image of the illumination chromaticity field and $\vec{\tau}$ is the vectorized ground truth illumination chromaticity field, and $\vec{\lambda}$ is the best single illuminant estimate. The goal is to find the best $\vec{\lambda}$ minimizing Err. Although it would seem that this might require an iterative optimization, the $\vec{\lambda}$ minimizing the error in Eq. (9) can be found, in fact, by direct computation [4]

$$\lambda_k = \frac{\rho_k \cdot \gamma_k}{\rho_k \cdot \rho_k} \tag{10}$$

Error Heatmap

Traditional IE test datasets typically provide only the chromaticity of a single illuminant as the ground truth. MIST instead provides pixel-wise ground truth, which allows us to plot a pixel-wise heatmap of the error in the illuminant estimate.

Results

Figure 3 shows the illumination estimates and error heat maps for the methods tested. Table 1 reports the mean, median and standard deviation of the three error measures for the methods tested. Clearly, none of the methods performs very well. In particular, the multi-illuminant methods are no better than the single-illuminant methods. The performance of Mixed k-means, the best multi-illuminant method, is not as good as the standard Grey Edge method, and neither is very close to the single-illuminant Oracle.

Conclusion

MIST, a new multi-illuminant, full-spectrum, fully ray-traced, synthetic image dataset has been presented and the performance of ten different illumination methods has been evaluated on it. Surprisingly, none of the methods specifically designed for spatially varying, multi-illuminant images results in a lower error than the ideal single-illuminant estimate method. The MIST dataset is freely available for download in order to encourage further research and testing in this area.

	rg-Distance Error			*Angular Error*			*Reproduction Angular Error*		
Methods	*Mean*	*Median*	*STDEV*	*Mean*	*Median*	*STDEV*	*Mean*	*Median*	*STDEV*
Do Nothing	0.12	0.12	0.05	17.95	17.77	8.49	16.52	17.28	6.87
Oracle	0.07	0.07	0.03	11.52	9.63	6.48	11.58	10.65	5.67
Max-RGB	0.13	0.11	0.07	17.35	16.66	8.70	17.56	17.07	7.98
Greyworld	0.11	0.10	0.06	15.18	13.68	8.20	15.13	14.17	7.15
Shades of Grey	0.12	0.10	0.07	16.32	15.63	8.62	16.42	16.06	7.80
Grey Edge	0.09	0.08	0.03	11.74	10.88	5.24	13.05	12.37	5.74
Weighted Grey Edge	0.11	0.10	0.04	14.48	13.62	6.39	14.89	14.42	6.36
FC4	0.09	0.09	0.03	12.39	11.46	5.56	13.23	12.52	5.87
Mixed K-Means	0.11	0.10	0.05	16.07	15.34	6.61	16.15	15.89	5.61
GAN(on subwindow)	0.12	0.13	0.05	14.41	14.01	5.93	16.78	15.87	7.42
Color by Colorization	0.12	0.14	0.06	17.62	17.10	7.23	16.27	15.71	6.59

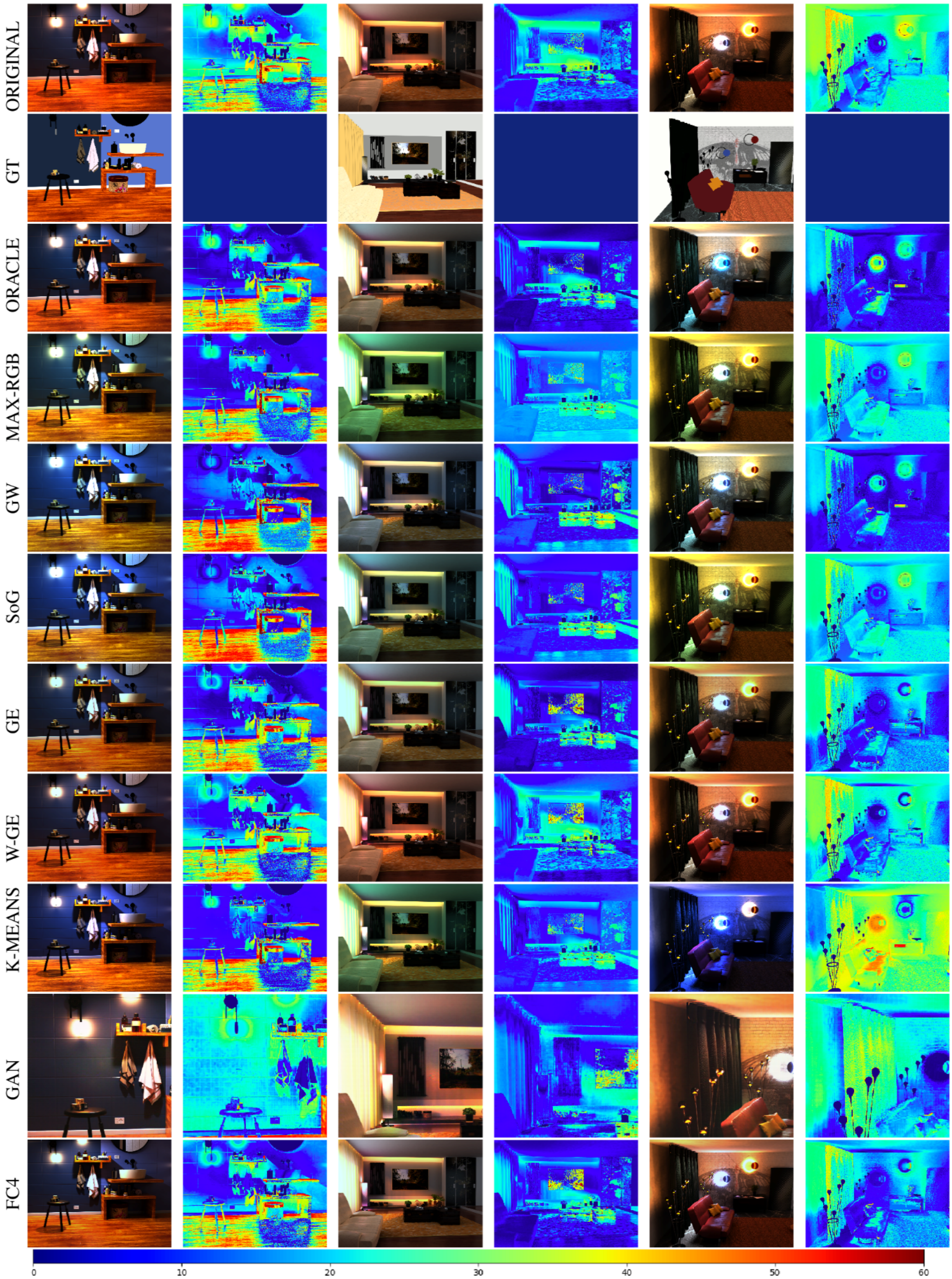

Figure 3. *Sample colour correction results for the different IE algorithms. The first row shows five different input images; the second row shows the corresponding ground truth reflectances imaged under D65; the third row shows colour correction based on the Oracle (best single-illuminant) result; and the subsequent rows present the results of nine different algorithms. Note that the GAN method only works with 512x512 square images, so it has been applied to a 512x512 subwindow of the corresponding input image. For display purposes, the images have been converted to standard non-linear sRGB.*

References

[1] S. Beigpour, M. Serra, J. van de Weijer, R. Benavente, M. Vanrell, O. Penacchio, D. Samaras, "Intrinsic image evaluation on synthetic complex scenes," in *IEEE International Conference on Image Processing*, 2013.

[2] S. Beigpour, ML. Ha, S. Kunz, A. Kolb, V. Blanz, "Multi-view Multi-illuminant Intrinsic Dataset," in *BMVC*, 2016.

[3] A.D. Logvinenko, B. Funt, H. Mirzaei, R. Tokunaga, "Rethinking Colour Constancy," *PLOS ONE DOI: 10.1371/journal.pone.0135029,* 2015.

[4] X. Hao and B. Funt , "A Multi-illuminant Synthetic Image Test," in *under review*, 2019.

[5] "Blender Foundation," [Online]. Available: https://www.blender.org/.

[6] M. Anderson, R. Motta, S. Chandraseka, "Proposal for a standard default color space for the internet—srgb," in Proc. Fourth *Color and Imaging Conference*, 1996.

[7] L. Xu, B. Funt, "How Multi-Illuminant Scenes Affect Illumination Estimation Performance," in *Proc. AIC 2015 International Colour Association Conference*, Tokyo, 2015.

[8] L. Shi, B. Funt, "Re-processed version of the Gehler color constancy dataset of 568 images," [Online]. Available: http://www.cs.sfu.ca/~colour/data/.

[9] P.V. Gehler, C. Rother, A. Blake, T. Minka and T. Sharp, "Bayesian Color Constancy Revisited," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[10] 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, 2017.

[11] A. Gijsenij, T. Gevers and J. van de Weijer, "Computational Color Constancy: Survey and Experiments," *IEEE Transactions on Image Processing,* vol. 20, no. 9, pp. 2475-2489, 2011.

[12] G.D. Finlayson, E Trezzi, "Shades of Gray and Colour Constancy," in *Twelfth Color Imaging Conference*, 2004.

[13] M. A. Hussain and A. S. Akbari, "Color Constancy Algorithm for Mixed-Illuminant Scene Images," *IEEE Access,* vol. 6, pp. 8964-8976, 2018.

[14] P. Das, AS Baslamisli, Y Liu, S Karaoglu and T. Gevers , "Color Constancy by GANs: An Experimental Survey," 2018.

[15] O. Sidorov, "Conditional GANs for Multi-Illuminant Color Constancy: Revolution or Yet Another Approach?" 2018.

[16] P. Isola, J.Y. Zhu, T. Zhou and A.A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *CVPR*, 2017.

[17] L. Zhu, and B. Funt, "Colorizing Color Images," in *Proc. Human Vision and Electronic Imaging XXIII*, 2018.

[18] Y. Hu, B. Wang, S. Lin, "FC 4: Fully Convolutional Color Constancy with Confidence-weighted Pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] 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.

[20] A. Gijsenij, T. Gevers and J. Van de Weijer, "Improving Color Constancy by Photometric Edge Weighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 5, pp. 918-929, 2012.