Consensus-Driven Illuminant Estimation with GANs

Marco Buzzelli, Riccardo Riva, Simone Bianco, and Raimondo Schettini
Department of Informatics, Systems and Communication
University of Milano – Bicocca, Italy

ABSTRACT

We present a method for illuminant estimation that exploits a generative adversarial network architecture to generate a spatially-varying illuminant map. This map is then transformed by consensus into a global illuminant estimation, in the form of a single RGB triplet. To this end, different consensus strategies are designed and compared in this paper. The best solution won second place in the 2nd International Illumination Estimation Challenge, specifically for the indoor track.

Keywords: Illuminant estimation, color constancy, white balance, Generative Adversarial Networks, GAN.

1. INTRODUCTION

Illuminant estimation aims to estimate the color of the light which illuminates the scene depicted in a digital image. Illuminant estimation can be exploited to perform Automatic White Balancing (AWB), i.e. correcting the image with the objective of rendering the scene as if it was acquired under a desired reference illuminant. This estimation-correction pipeline is one of the possible approaches to achieve so-called computational color constancy, which tries to mimic the ability of the human visual system to perceive consistent object colors under different illumination conditions [1]. Several algorithms for AWB have been proposed through the years, starting from those based on the extraction of low-level image statistics: such algorithms were successfully embedded into a unique generalized mathematical framework by van de Weijer et al. [2]. Gamut mapping methods [3][4] were later developed, based on the rationale of taking image data captured under an unknown light to a gamut of reference colours taken under a known light. Significant improvements were then reached through so-called parametric solutions, which are typically based on handcrafted features and which depend on few manually-tunable parameters, the most effective to date being Akbarinia et al. [5] and Cheng et al. [6]. Since the explosion of Convolutional Neural Networks (CNNs) in computer vision, several works have been leveraging such computational models to approach the problem of illuminant estimation, obtaining significant results [7][8][9][10]. In recent years, Generative Adversarial Networks (GANs) have shown remarkable results in various computer-vision-related tasks [11][12], including color constancy [13].

In this paper, we exploit Generative Adversarial Networks to address the problem of illuminant estimation, under the framework defined in the 2nd International Illumination Estimation Challenge [14], specifically for the indoor track. In Section 2 we present our method, which produces a spatially-varying estimation that is then transformed into a global estimation through an appropriate consensus strategy. In Section 3 we define the experimental setup used to develop and test our solution, we present its performance in the framework of the challenge. We discuss some visual insights on the internal mechanics of our proposed solution, thus highlighting both strengths and weaknesses of the proposed method.

2. PROPOSED METHOD FOR ILLUMINANT ESTIMATION

We devise a spatially-varying illuminant estimation model, which is trained by exploiting the mechanisms of adversarial training, having as input the global illuminant information. In doing so, we intend to take advantage of the intrinsic ability of GANs to autonomously define a loss function, which corresponds to the discriminator module. In order to address the typical artifacts introduced by generative models, we investigate several consensus strategies to process the generated spatially-varying illuminant map into a single global illuminant estimate based on the von Kries model [15], which describes the correction of the tristimulus values (e.g. red, green and blue for a digital image) with a diagonal matrix.

2.1 Adversarial Training

The training process is illustrated in Figure 1. Let “Input image” be a three-channel color image, and let “Global illuminant” be the illuminant ground truth, assumed to be in the form of three coefficients according to the von Kries model. We compute the corresponding color-corrected image “Target image” by dividing each pixel of each channel of “Input image” by the corresponding channel coefficient in “Global illuminant”. The image pair “Input image”-“Target image” is used to train a Generative Adversarial Network model based on the standard pix2pix architecture [16].

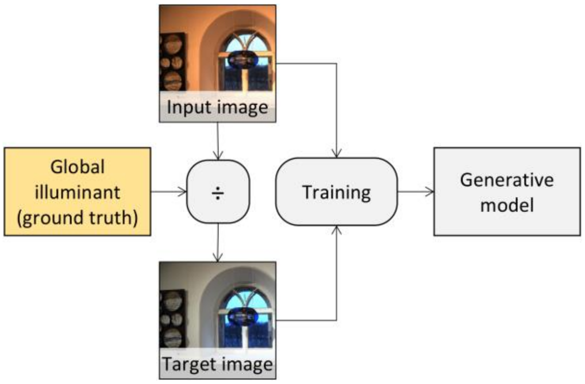


Figure 1. Schematic representation of the data involved in the training of our generative model

The pix2pix is a conditional Generative Adversarial Network model developed by Isola et al. [16]: the core idea of their model is to have a generator network and a discriminator network, to be trained in an adversarial fashion in order to make them progressively more precise at their task. The architecture of the generator network is based on the U-net model [17]: an autoencoder-like network [18] with skip connections [19]. The first part of the network has the purpose of compressing the input data (originally set to 224×224 pixels) by repeating standard blocks (convolution / batch normalization / leaky ReLU) until it reaches a bottleneck layer, where the dimensions of the activation reduce to a compact mono-dimensional vector (with cardinality 512 in our implementation). Subsequently, a decompression section is implemented with repeated blocks (transposed convolution / batch normalization / dropout) alternated to concatenations, in order to implement skip connections that have the same activation shape from the compression section. The discriminator network is a Convolutional Neural Network that starts with a concatenation layer in order to merge together the two images provided as input, i.e. the original input image and the generated image. The output of this discriminator network is a dense 16×16 activation map. Through the PatchGan architecture [16], in fact, each output value describes the likelihood of the corresponding patch in the input image (70×70 pixels) to be real, as opposed to be fake, i.e. generated. The training procedure is performed by having the discriminator alternately receive fake inputs as output by the generator, and real inputs from the dataset. We refer to Isola et al. [16] for more details.

The network is trained with batch size equal to one due to virtual memory occupation limits, for a total of 100 epochs determined on the basis of an inspection of preliminary results.

2.2 Consensus-based Inference

The inference process is illustrated in Figure 2. At inference time, the trained generative model is used to produce the “Generated image”, that is an initial estimate of the color-corrected version of the input image. This intermediate output implies a per-pixel color correction. In fact, dividing pixel-by-pixel “Input image” by “Generated image” we obtain a spatially-varying illuminant estimation for each pixel that is independent from all the others, and which is a punctual von Kries transformation.

Assuming a single uniform illuminant, we need a consensus strategy to effectively map this spatially-varying internal representation into a unique global illuminant. This step makes it also possible to handle any artifact potentially introduced by the generative model, such as grid-like patterns and sparse regions of pixels with extremely strong chromaticities (from now on referred to as “hot pixels” due to their similarity with the corresponding phenomenon in digital photography). Visual examples of such artifacts will be shown in the experimental section. By reprocessing the input image with a global illuminant elected by consensus over the spatially-varying generated one, we can successfully remove the aforementioned artifacts from the final output.

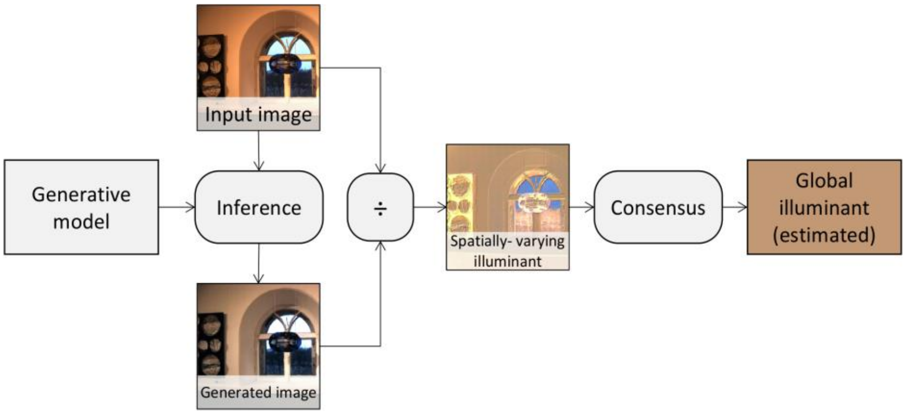


Figure 2. Schematic representation of the inference process. The global illuminant estimation is defined as the consensus over a spatially-varying illuminant coming from our generative model.

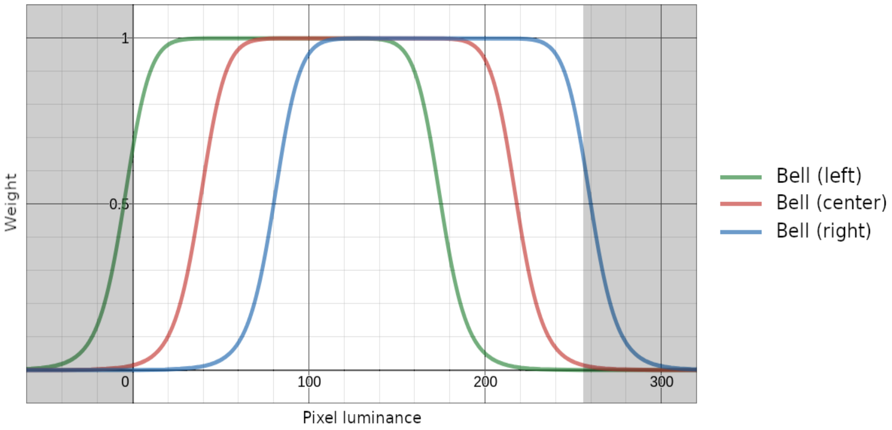


Figure 3. Bell-shaped weighting functions used to average estimated saliency based on pixel luminance.

We explore different alternative strategies to elect a consensus from the estimated “Spatially-varying illuminant” map:

- The baseline strategy is the *mean* of each color channel, considered independently.
- In order to exclude outliers in the distribution of estimated illuminants, due for example to the aforementioned hot pixels, the *median* of each channel is used as an alternative consensus strategy.
- A third strategy involves assigning different weights to the estimated spatially-varying illuminants, and consequently performing a *weighted mean*. The weights are computed as a function of the luminance of the input image using a bell-shaped function *B(x)* centered at three different luminance levels *c*:

$$B(x) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}} \qquad (1)$$

Given Equation 1, with *a* = 90, *b* = 6, and *c* set to 255 ·⅓, 255 ·½, and 255 · ⅔, we obtain the three weighting functions shown in Figure 3, respectively labelled “Bell (left)”, “Bell (center)” and “Bell (right)”. A weighting function over pixel luminance has the effect of reducing the impact of low-intensity and high-intensity pixels in the estimated global illuminant to be elected by mean consensus. Specifically, when the spatially-varying illuminant is computed as the ratio between input image and generated image, low-luminance pixels are subject to numerically-unstable results, and the corresponding illuminants are therefore considered unreliable. On the other hand, highluminance pixels are potentially affected by non-linearities of the sensor close to its saturation level (including completely clipped pixels), and as such these should also be considered less reliable sources of information.

- Finally, a meta strategy is developed with the aim of dynamically predicting for each image the most appropriate consensus strategy among the ones just defined above. To this end, a VGG-16 neural network [20] has been trained to address the task as a five-class classification problem.

3. EXPERIMENTAL RESULTS

3.1 Experimental Setup

The presented method was developed for the 2nd International Illuminant Estimation Challenge. The challenge is composed of three tracks: general, indoor, and two-illuminant. All images in the challenge dataset were taken with either a Canon 550D camera or with a Canon 600D camera. In the lower-right corner of each image, the SpyderCube calibration target was placed. The two 18% gray faces facing the camera were used to determine the ground truth illuminant for each image, expressed according to the von Kries model. As a pre-processing step common to all images, the camera black level (equal to 2048) has to be subtracted from the 16-bit-encoded files.

The “indoor” track was selected as the objective of the experiments presented in this paper, in order to define a constrained environment for the assessment of the developed solution. The training set of the indoor track is composed of 658 images. These are only images with an angular difference less than 2 degrees between left and right faces of the SpyderCube, and the brightest of the two faces is indicated as the reference ground truth. The test set of the indoor track is composed by 58 unseen images.

In addition to the official training set, the existing Cube+ dataset [21] was made available as bonus data. The Cube+, distributed as an extension of the Cube dataset, contains a total of 1707 images, shot with the same Canon 550D camera that was used in the challenge data acquisition. Also for the Cube+ dataset, the ground truth data is extracted using a SpyderCube included in the scene.

We evaluate our experiments using both the recovery angular error err_{rec} [22] and the reproduction angular error err_{rep} [23], computed between a ground truth illuminant $U = (u_R, u_G, u_B)$ and an estimated illuminant $V = (v_R, v_G, v_B)$:

$$err_{rec} = \arccos\left(\frac{U \cdot V}{|U||V|}\right) = \arccos\left(\frac{\sum_i u_i v_i}{\sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2}}\right) \tag{2}$$

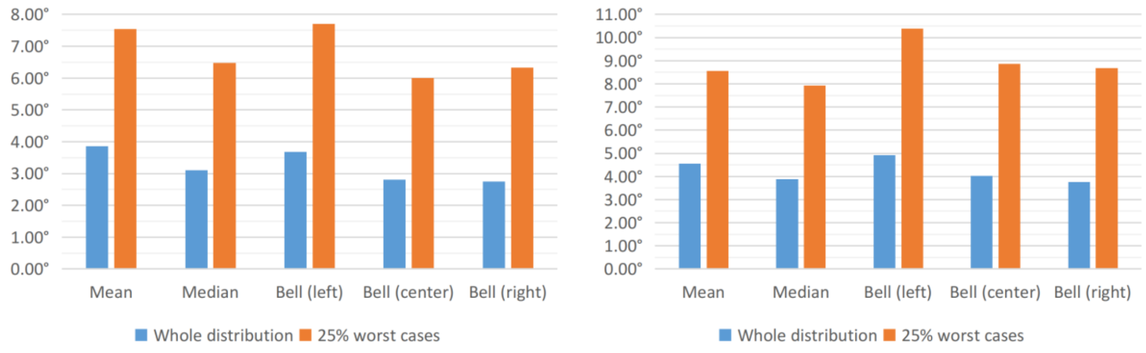
$$err_{rep} = \arccos\left(\frac{\frac{U}{V} \cdot (1,1,1)}{|\frac{U}{V}|\sqrt{3}}\right) = \arccos\left(\frac{\sum_i \frac{u_i}{v_i}}{\sqrt{\sum_i \frac{u_i^2}{v_i^2}}\sqrt{3}}\right) \tag{3}$$

We develop and investigate our solutions by training on the Cube+ dataset and testing on the training set of the indoor track. Our final solution, submitted to the competition, was eventually trained on the Cube+ dataset, and fine-tuned on the training set of the indoor track.

3.2 Quantitative Assessment

Figure 4 reports the results for the developed consensus strategies, trained on the Cube+ dataset, and tested on the training set of the indoor track. For each experiment and error metric, we report the average error of the whole test error distribution, as well as the average error of the 25% worst results. Focusing on worst-result statistics, in fact, provides a valuable piece of information that describes the ability of a given method to handle extreme and special cases.

Concerning the recovery error, the best-performing strategy appears to be the application of either a center-localized or right-localized Bell weighting function, which produce comparable results. The second best strategy is the median, which effectively removes outliers from the generated spatially-varying illuminant map. The analysis based on reproduction error shows similar results in general. For the 25% worst cases the best strategy appears is the application of the median. As the illuminant estimation challenge is evaluated on reproduction error, we focused our attention particularly on this configuration. Furthermore, we consider the “25% worst cases” metric to be a robust indicator of the performance of a given method in unknown/unpredictable conditions.



(a) Recovery error

(b) Reproduction error

Figure 4. Results of our illuminant estimation solution based on different consensus strategies, in terms of (a) recovery error and (b) reproduction error. The reported statistics refer to a model trained on the Cube+ dataset and tested on the training set of the 2nd illuminant estimation challenge (indoor track).

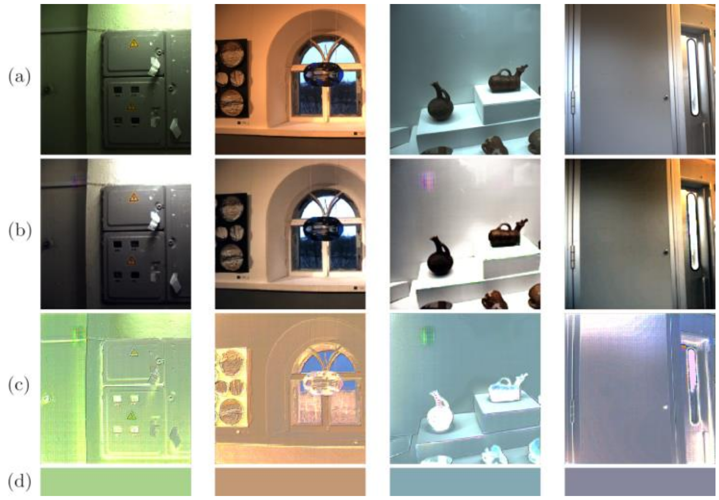


Figure 5. Visual inspection of our proposed consensus-based illuminant estimation model. (a) input image, (b) corrected image as output from the generative model, (c) spatially-varying illuminant estimation, (d) consensus-based global illuminant estimation.

For the aforementioned reasons, we eventually adopted the median-based consensus strategy in our final challenge entry (labelled MCGAN as in “Median-Consensus GAN”). The meta strategy designed to dynamically select the best consensus strategy was also submitted to the challenge (labelled PCGAN as in “Predicted-Consensus GAN”), however, it was found to collapse to the median selection in 53 out of 58 images, thus resulting in comparable performance to MCGAN. Table 1 reports the official evaluation on the challenge test set, where our MCGAN entry won second place.

Table 1. Reproduction error from the official leaderboard of the 2nd illuminant estimation challenge (indoor track)

Algorithm	**Mean**	**Median**	**Trimean**
Constant (baseline)	15.270 °	14.802 °	15.332 °
PCGAN (ours)	3.376 °	2.312 °	2.337 °
MCGAN (ours)	3.301 °	2.312 °	2.298 °
sde-awb	2.541 °	1.763 °	1.943 °

3.3 Visual Inspection

In Figure 5 we report a visual inspection over the outputs of our model for illuminant estimation on some sample images taken from the challenge dataset. For each reported example, row (a) shows the input image, row (b) shows the image generated by the generative adversarial model, row (c) shows the pixel-by-pixel ratio between rows (a) and (b), effectively resulting in a spatially-varying illuminant. Row (d) shows the final estimated triplet based on median consensus, which is the strategy adopted in our entry to the competition. For visualization purposes, we have processed all the images by setting an arbitrary fixed white point that removes the “greenish” dominant of the RAW images.

The second-column image shows that multiple sources of light can be handled by the proposed model, which in fact produces a yellowish estimate for the interiors, and a blueish estimate for the scene outside the window. Conversely, the last column shows another multiple illuminant image where our model incorrectly generates a close-to-uniform illuminant estimation. The first and third columns showcase examples of the generative model producing artifacts (the magenta spot in the top left corner), which are correctly filtered-out with the help of the appropriately-devised consensus strategy.

4. CONCLUSION

We have presented a method for illuminant estimation that won second place in the indoor track of the 2nd Illuminant Estimation Challenge. Our solution is based on the estimation of a spatially-varying illuminant through adversarial training, which is then translated into a global illuminant through the application of a simple yet effective consensus strategy.

This work constitutes the basis for several further developments. The single-illuminant scenario which is the subject of this paper can be further explored through the definition of more advanced consensus strategies. Considering a multiple-illuminant scenario, the intermediate representation generated by the adversarial model could be exploited as a starting point for illuminant clustering.

REFERENCES

[1] Foster, D. H., “Does colour constancy exist?,” Trends in cognitive sciences **7**(10), 439–443 (2003).

[2] Van De Weijer, J., Gevers, T., and Gijsenij, A., “Edge-based color constancy,” IEEE Transactions on image processing **16**(9), 2207–2214 (2007).

[3] Forsyth, D. A., “A novel algorithm for color constancy,” Int. Journal of Computer Vision **5**(1), 5–35 (1990).

[4] Finlayson, G. D., Hordley, S. D., and Tastl, I., “Gamut constrained illuminant estimation,” International Journal of Computer Vision **67**(1), 93–109 (2006).

[5] Akbarinia, A. and Parraga, C. A., “Colour constancy beyond the classical receptive field,” IEEE transactions on pattern analysis and machine intelligence **40**(9), 2081–2094 (2017).

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

[7] Bianco, S., Cusano, C., and Schettini, R., “Color constancy using cnns,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 81–89 (2015).

[8] Bianco, S., Cusano, C., and Schettini, R., “Single and multiple illuminant estimation using convolutional neural networks,” IEEE Transactions on Image Processing **26**(9), 4347–4362 (2017).

[9] Buzzelli, M., van de Weijer, J., and Schettini, R., “Learning illuminant estimation from object recognition,” 2018 25th IEEE International Conference on Image Processing (ICIP), 3234–3238, IEEE (2018).

[10] Bianco, S. and Cusano, C., “Quasi-unsupervised color constancy,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 12212–12221 (2019).

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., “Generative adversarial nets,” Advances in neural information processing systems, 2672–2680 (2014).

[12] Wang, Z., She, Q., and Ward, T. E., "Generative adversarial networks in computer vision: A survey and taxonomy," arXiv preprint arXiv:1906.01529 (2019).

[13] Das, P., Baslamisli, A. S., Liu, Y., Karaoglu, S., and Gevers, T., "Color constancy by gans: An experimental survey," arXiv preprint arXiv:1812.03085 (2018).

[14] Ershov, E., Savchik, A., Semenkov, I., Terekhin, A., Senshina, D., Nikolaev, D., Banic, N., Subasic, M., and Loncaric, S., "2nd international illumination estimation challenge," (2020). http://chromaticity.iitp.ru/ (accessed September 14, 2020).

[15] von Kries, J., "Theoretische studien ¨uber die umstimmung des sehorgans," Festschrift der Albrecht-LudwigsUniversit ¨at -, 145–158 (1902).

[16] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," Proceedings of the IEEE conference on computer vision and pattern recognition, 1125–1134 (2017).

[17] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," Int. Conference on Medical image computing and computer-assisted intervention, 234–241, Springer (2015).

[18] Kramer, M. A., "Nonlinear principal component analysis using autoassociative neural networks," AIChE journal **37**(2), 233–243 (1991).

[19] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778 (2016).

[20] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).

[21] Bani´c, N., Koˇsˇcevi´c, K., and Lonˇcari´c, S., "Unsupervised learning for color constancy," arXiv preprint arXiv:1712.00436 (2017).

[22] Hordley, S. D. and Finlayson, G. D., "Reevaluation of color constancy algorithm performance," JOSA A **23**(5), 1008–1020 (2006).

[23] Finlayson, G. D. and Zakizadeh, R., "Reproduction angular error: An improved performance metric for illuminant estimation," perception **310**(1), 1–26 (2014).