






Analysis of biases in automatic white balance datasets and methods

Marco Buzzelli¹  | Simone Zini¹  | Simone Bianco¹  | Gianluigi Ciocca¹  |
Raimondo Schettini¹  | Mikhail K. Tchobanou²

¹Department of Informatics Systems and Communication, University of Milano-Bicocca, Milan, Italy

²Moscow Research Center, Huawei Technologies Co. Ltd, Moscow, Russian Federation

Correspondence

Marco Buzzelli, University of Milano - Bicocca, Department of Informatics Systems and Communication, Milan 20126, Italy.

Email: marco.buzzelli@unimib.it

Abstract

Annotated datasets for automatic white balance (AWB) are used for the evaluation and, when necessary, the training, of AWB methods. Relying on such datasets requires awareness of the potential bias in their content and characteristics: some methods are designed to rely on the presence of particular elements, such as human skin, while other methods learn implicit relationships between image content and light properties from training data. The dependency on these relationships makes it fundamental to understand whether the available datasets are actually representative of common application scenarios, such as the presence of human subjects, the diversity of composition, or the illumination conditions. In this paper we overview the most common datasets for Automatic White Balance, including those for single as well as multiple illuminant estimation, providing a critical analysis on their characteristics. Furthermore, we identify a number of existing methods for single illuminant estimation, as a representative pool of approaches to the problem with various levels of complexity. We investigate how the performance of these correlate to the image content of common datasets.

KEYWORDS

automatic white balance, color constancy, dataset analysis, illuminant estimation

JEL CLASSIFICATION

C8: Data Collection and Data Estimation Methodology • Computer Programs

1 | INTRODUCTION

The advent of data-hungry learning approaches to solve various computer vision tasks has highlighted their dependency on the datasets used for training and verification. The community of color imaging is not exempt from this phenomenon, in particular the field of automatic white balance (AWB), which is the task of

correcting a digital image as if the scene was captured under some reference illumination conditions. AWB aims at replicating in the digital domain a mechanism of the human visual system called “color constancy,” which allows people to stably perceive the chromatic property of objects, completely or partially disregarding the source of light.¹ As such, Automatic White Balance can also be referred to as computational color constancy.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. Color Research and Application published by Wiley Periodicals LLC.

The most effective^{2,3} recent methods for AWB, often based on machine learning and in particular on convolutional neural networks, are trained and validated on specialized datasets. However, in some cases, these datasets are characterized by limited diversity of content. In this work, we analyze the most commonly used datasets for AWB from different points of view: distribution of the illuminants, image acquisition parameters, and semantic content of the considered scenes. This analysis aims at raising awareness about the extent to which such datasets are representative of common application cases. A number of studies in the literature tried to describe and model the distribution of image content in photo collections. Hultgren et al⁴ provided in 2008 a data survey based on the concept of “photospace,” where they analyzed the distribution of about 500 user photos in terms of subject distance to the camera and illumination levels, observing a high concentration of short-distance subjects in relatively low illumination conditions. Hudelist et al⁵ in 2015 conducted a user survey among photographers, to investigate the most popular motivations for shooting. The resulting top-most motives included immortalizing human subjects in various events as well as long-range landscapes. Ferwerda et al⁶ in 2018 inspected the pictures published in the Instagram accounts of about 200 participants to their study. Although the precise distribution of image content was not disclosed, commonly identified categories included architecture and close-up views (body parts, clothing items, etc.) A research led for the Honor smartphone line⁷ in 2019 delved into the trends of smartphone use for photo acquisition by European users, showing that the average subject in the study sample takes about 600 selfies per year. Additional insights in the imaged content also suggested a tendency to prefer outdoor in-the-wild scenarios. These studies can be used as a broad indication of what type of content is most common in personal photo collections, and thus they can suggest to what extent a given dataset is representative for a real-world application, therefore focusing on the distance of depicted subjects, illumination levels, presence of human subjects, distinction between close-up views and indoor/outdoor environments, and general composition of landscape shots.

The fact that AWB datasets are sometimes used in international challenges, and the growing tendency in the scientific community to evaluate the quality of a method only in quantitative terms, have polarized the authors to look for marginal performance improvements, sometimes in the order of few decimal degrees in terms of angular error. The practical impact on human perception of such small variations, however, has been widely studied and debated.⁸⁻¹³ Regarding the content of the image, then, we know that this is increasingly used by methods of Automatic White Balance, both in explicit

form,^{14,15} and implicitly within neural network learned features.^{16,17} Bianco et al,¹⁵ for example, proposed a fully automatic method to exploit the skin color extracted from detected faces to estimate the illuminant in the scene. While it has been shown that personal photo collections often include human faces, we will show that this is not true for AWB datasets. Buzzelli et al¹⁷ proposed a learning strategy that requires no explicit AWB annotations, exploiting instead object-recognition labels, where a classifier is used as a proxy loss function for illuminant estimation. This approach showed that the semantics of the imaged content do indeed contain some information that is beneficial to AWB. In light of these considerations, the analysis presented in this paper can be useful not only to design and acquire datasets more representative of reality, but also to provide an incentive to the adoption of more insightful evaluation protocols for AWB methods, for example inspecting their performance on specific classes of content.

The analysis presented in this paper extends our preliminary work¹⁸ by introducing considerations about multi-illuminant datasets, and by observing the relationship between the image content and the performance of various AWB methods. The document is organized as follows: Section 2 describes the most common datasets used in AWB algorithms training and evaluation with focus on the image content, and illuminants variability. Both single-illuminant and multiple-illuminant datasets are considered. In Section 3, we analyze the intrinsic biases of the aforementioned datasets. In Section 4, the performance of selected single-illuminant AWB algorithms are analyzed with respect to the datasets content. Section 5 concludes the paper by summarizing our findings and providing some suggestions for future directions.

2 | ANALYZED DATASETS

2.1 | Single-illuminant datasets

There are several single-illuminant datasets available in the literature. Here we focus our attention to the most widely used for benchmarking Automatic White Balance algorithms. The considered datasets are: ColorChecker by Gehler et al¹⁹; Cube++ and SimpleCube++ by Ershov et al²⁰; Gray Ball by Ciurea et al²¹; INTEL-TAU by Laakom et al²²; and NUS by Cheng et al.²³ In the following, we briefly describe these datasets, whose main characteristics are summarized in Table 1.

- *Color Checker*: Initially proposed by Gehler et al,¹⁹ the dataset is composed of 568 images of indoor and outdoor scenes. Eighty-six images were shot with a Canon EOS-1DS, and the remaining 482 with a Canon EOS

TABLE 1 Analyzed datasets for single-illuminant Automatic White Balance, with their main characteristics

Name	References	Year	Cameras	Images	Reference target
ColorChecker	Gehler et al ¹⁹	2008	2	568	24-patch Macbeth Color Checker
Cube++	Ershov et al ²⁰	2020	2	4890	Datacolor SpyderCUBE
Gray Ball	Ciurea et al ²¹	2003	1	11 346	Gray sphere
INTEL-TAU	Laakom et al ²²	2021	3	7022	X-Rite ColorChecker Passport chart
NUS	Cheng et al ²³	2014	9	1853	24-patch Macbeth Color Checker

5D. A copy of the “Macbeth” color chart is included in each scene and used to annotate the image illuminant. It is one of the most widely used dataset and, through the years, changes to the ground truth have been proposed, such as the “reprocessed” version by Shi and Funt,²⁴ and the more recent “recommended” version by Hemrit et al.²⁵ For the purpose of this work, we used this latest version.

- *Cube++ and SimpleCube++*: The Cube++ dataset is an iteration of the previous “Cube” and “Cube+.” It contains a total 4890 images: 2361 shot using a Canon EOS 550D camera, and 2529 with a Canon EOS 600D. Its two neutral 18% gray faces were used to determine the ground-truth illumination for each image, thus potentially providing two annotations per image. A subset of this dataset, named SimpleCube++, is composed of 2234 images that have less than one degrees difference between left and right ground truth illumination estimation. This is the version used for the analyses in this paper.
- *Gray Ball*: This dataset consists of 11 346 near-consecutive images extracted from 15 video sequences recorded with a video camera. The sequences were taken in various indoor and outdoor scenes. Many of the images also depict people. The dataset was collected using a Sony VX-2000 digital video camera. In the bottom-right corner of each image is present the eponymous gray ball color target used for illuminant annotation. The images are provided in non-linear 8bit RGB format, which is found to hinder the performance of many AWB algorithms.
- *INTEL-TAU*: This dataset is currently the largest available AWB dataset. It contains 7022 images captured using three different camera models: Canon 5DSR, Nikon D810, and Sony IMX135. The dataset has mainly real scenes and some lab printouts. It is a revised version of the INTEL-TUT image dataset²⁶ that was released in 2017 and later withdrawn due to privacy concerns. In the current dataset all recognizable faces, license plates, and other privacy sensitive information have been masked. Ground truth annotations have been collected and associated to image batches using a X-Rite ColorChecker Passport chart, positioned as to reflect the main illumination source.

- *NUS*: The dataset is a collection of 1853 images collected by the National University of Singapore (NUS). Nine different cameras has been used to acquire the images: Canon EOS-1Ds Mark III (259 images), Canon EOS 600D (200 images), Fujifilm X-M1 (196 images), Nikon D40 (117 images, not included in the more common “NUS-8” version of the dame dataset) Nikon D5200 (200 images), Olympus E-PL6 (208 images), Panasonic Lumix DMC-GX1 (203 images), Samsung NX2000 (202 images), and Sony SLT-A57 (268 images). All images include a 24-patch Macbeth Color Checker target, and depict indoor, outdoor, close-up scenes, as well as people.

2.2 | Multi-illuminant datasets

The aforementioned datasets for Automatic White Balance rely on the assumption of a single global illuminant present in the scene. This simplification facilitates research in the challenging domain of global illuminant estimation. However, it neglects the reality that most real-life scenes are in practice affected by multiple illuminants, possibly due to different light sources at different correlated color temperatures, mutual surface inter-reflections, or the coexistence of sun and shadow areas. To this extent, a growing field of research has focused on the generation and exploitation of datasets for Automatic White Balance in multiple-illuminant scenarios. One intrinsic difficulty of this approach lies in accurately capturing spatially-varying illuminant information for ground truth generation. Among the available datasets, the most notable ones are multi-illuminant dataset (MID) by Bleier et al²⁷; multi-illuminant multi-object (MIMO) by Beigpour et al²⁸; multi-illuminant synthetic test set (MIST) by Hao et al²⁹ and Drone by Aghaei et al.³⁰ Table 2 summarizes the main characteristics of these datasets.

- *MID*: The dataset depicts four laboratory scenes (“chalk,” “figures,” “fruits,” “rabbit”), illuminated alternatively by two Reuter’s lamps and ambient light, and captured with a Canon EOS 550D (Sigma 17-70 lens) for a total of 36 combinations. Different versions

Name	References	Year	Images	Scenes	Ground truth
MID	Bleier et al ²⁷	2011	420	4	Per patch
MIMO	Beigpour et al ²⁸	2014	78	10 lab +20 real-world	Per pixel
MIST	Hao et al ²⁹	2020	900	3 × 50 view points	Per pixel
DRONE	Aghaei et al ³⁰	2020	31	31	Sparse locations

TABLE 2 Multi-illuminant datasets for Automatic White Balance, with their main characteristics

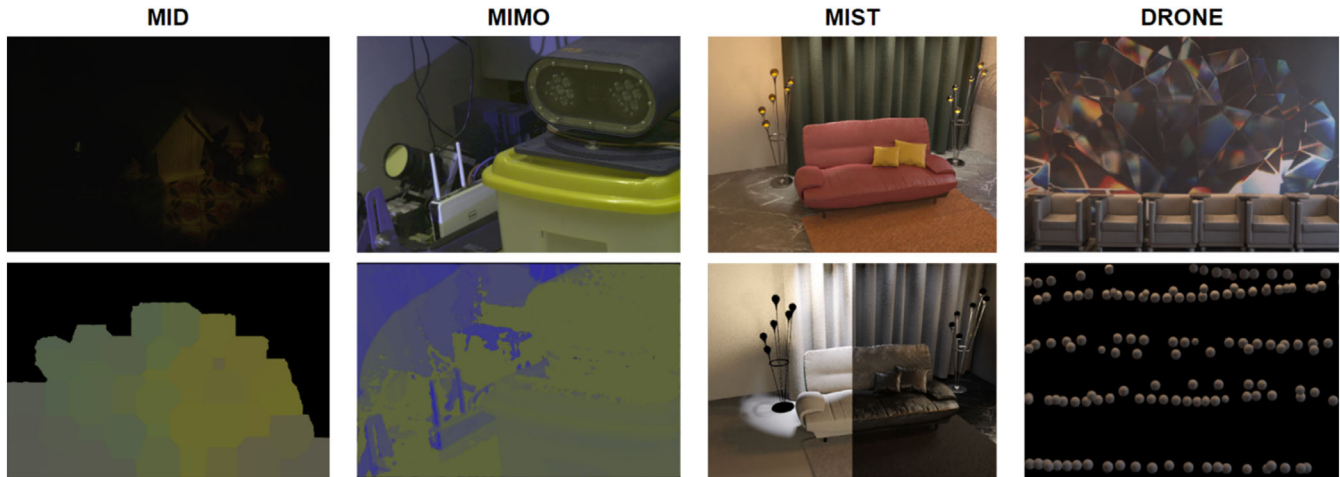


FIGURE 1 Image-ground truth pairs for different multi-illuminant datasets. MID provides per-patch information. MIMO per pixel. MIST per pixel, including both diffuse and specular reflection (image from the corresponding paper). DRONE sparse grayball-based information (image from the corresponding paper). MID, multi-illuminant dataset; MIMO, multi-illuminant multi-object; MIST, multi-illuminant synthetic test set

of each setup are made available, for example by varying the reflective properties of the objects and the camera exposure, for a total 420 images. The ground truth information is provided in non-uniform patches, as shown in Figure 1, and it was collected by spray-painting the captured scenes with grey paint. The authors also note that this method of ground-truth generation eliminates information about objects inter-reflections.

- *MIMO*: This dataset depicts scenes in two setups: a laboratory part set indoor with controlled illumination settings, and a real-world part in less-constrained environments. All the images were collected using a Sigma SD10 camera based on a Foveon X3 sensor. The laboratory images in particular were exposed to “red,” “blue,” and “white” illuminants, and each scene was also captured under only a single illuminant from each position to generate the per-pixel ground truth. The real-world images were exposed to either one indoor/outdoor light and a colored projector, or to sun and no-sun (shadow/ambient light). As such, the ambient illuminant is present on almost the entire image area, while the direct illuminant covers only a part of each scene.
- *MIST*: The dataset overcomes the practical limitations of dense ground truth collection, by relying on

synthetically generated scenes. It was designed for use beyond white balance, therefore each rendered scene is associated to a rich set of per-pixel information: percent surface spectral reflectance, spectrum of the incident illumination, separate specular and diffuse components of the reflected light, and depth. The total 900 images stem from three scenes rendered from 50 view points and in six different illuminant conditions, using an extension of the Blender Cycles ray-tracing renderer.

- *DRONE*: This dataset presents a wide variety of indoor and outdoor real-world scenarios, for a total of 31 scenes. The images are captured using a Nikon D700 DSLR (Nikon 50 mm 1:1.4G lens). The peculiarity of the dataset is that the problem of spatially varying ground truth collection is captured by flying a drone mounted with a uniform gray plastic ball, whose reflectance properties are measured using the Photo Research SpectraScan PR650 spectroradiometer. Each scene is accompanied by an average of 100 shots with the drone in different positions, and a composite sparse patch-wise ground truth is generated based on these shots.

In addition to the aforementioned multi-illuminant specific datasets, Cheng et al³¹ identified a subset of

66 images in the ColorChecker dataset, and 34 images from the non-AWB RAISE dataset,³² as being illuminated by two light sources. Furthermore, the single-illuminant Cube++ dataset is acquired with a SpyderCUBE color target, which allows capturing incident light information from different sources. Some images within the dataset reach an angular difference of 10°, which suggests the presence of distinct light sources. However, no such information is directly associated to specific image regions. Images from the same camera and color target were also used for the “Two-illuminant track” of the 2nd International Illumination Estimation Challenge.³³ Also in this case the two illuminant sources were provided without spatial reference, and the squared sum of two angular reproduction errors was used for illuminant estimation assessment.

3 | BIASES INTRINSIC IN AWB DATASETS

In this section, we analyze the intrinsic biases of the available datasets for Automatic White Balance, in terms of distribution of the illuminants, image content, shooting parameters and illumination levels.

3.1 | Illuminants distribution

3.1.1 | Single-illuminant datasets

Most existing datasets for Automatic White Balance are associated with ground truth information in the form of Red, Green, and Blue (RGB) triplets, defined in the camera-specific RAW RGB color space. This type of information implies the application of the estimated illuminant for white balance based on a von Kries-like transform,³⁴ using a diagonal matrix with multipliers a , b , c to apply independent correction to, ideally, the response of cone photoreceptors, or to the three color channels:

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (1)$$

This practice, although commonly adopted, is known to be sub-optimal and unable to fully handle metameric effects,³⁵ and it suggests the need for richer ground truth annotations in the collection of future datasets. Nonetheless, it is the balance technique of choice when handling this type of information.

TABLE 3 Handpicked sunlight values (D55-like) for the different cameras of popular single-illuminant AWB datasets

Dataset	Camera	R/G	B/G
ColorChecker	Canon EOS-1DS	0.5990	0.7969
	Canon EOS 5D	0.5072	0.6791
SimpleCube++	Canon EOS 550D	0.4486	0.6739
	Canon EOS 600D	0.4456	0.6807
Gray Ball	Sony VX-2000	1.0000	1.0000
INTEL-TAU	Canon 5DSR	0.4366	0.5996
	Nikon D810	0.5003	0.7610
	Sony IMX135	0.5650	0.6436
NUS	Canon 1Ds MkIII	0.4832	0.6867
	Canon 600D	0.4543	0.6716
	Fujifilm XM1	0.5200	0.6015
	Nikon D40	0.4551	0.6814
	Nikon D5200	0.3967	0.6455
	Olympus EPL6	0.5047	0.5592
	Panasonic GX1	0.4223	0.6630
	Samsung NX2000	0.3981	0.6382
Sony A57	0.3578	0.6651	

Based on the observation that each dataset is annotated with RGB data in its own camera-specific RAW RGB space (derived from the spectral sensitivities of the sensors involved in the acquisition), a direct comparison of the data distributions is not possible. Ideally, to do so we would need to map the datasets illuminants into a device-independent color space, taking into account said spectral sensitivities. These, however, are often not available. An alternative option is therefore to normalize the datasets with respect to a reference white, selected from manual inspection of the images and formulated in the form of a RAW RGB triplet, and applied by referring to Equation (1). Despite the aforementioned limitations of this approach, it allows for a certain degree of uniformity across different data sources. More specifically, we chose direct sunlight as the reference white, since images in this setup could be easily identified by observing the sky and/or cast shadows. By referring to a definition from the Encyclopedia of Color Science and Technology,³⁶ “Illuminant D55 can be assumed to represent the SPD (Spectral Power Distribution) for (direct) sunlight provided that the sun is not too low in the sky.” Then, for each camera, we handpicked an image that best-represents the previous definition. The image was selected by also considering metadata such as timestamp and approximate location. We thus used the annotation of the selected images as the corresponding camera reference white in RAW-specific RGB form. The chromaticities of the

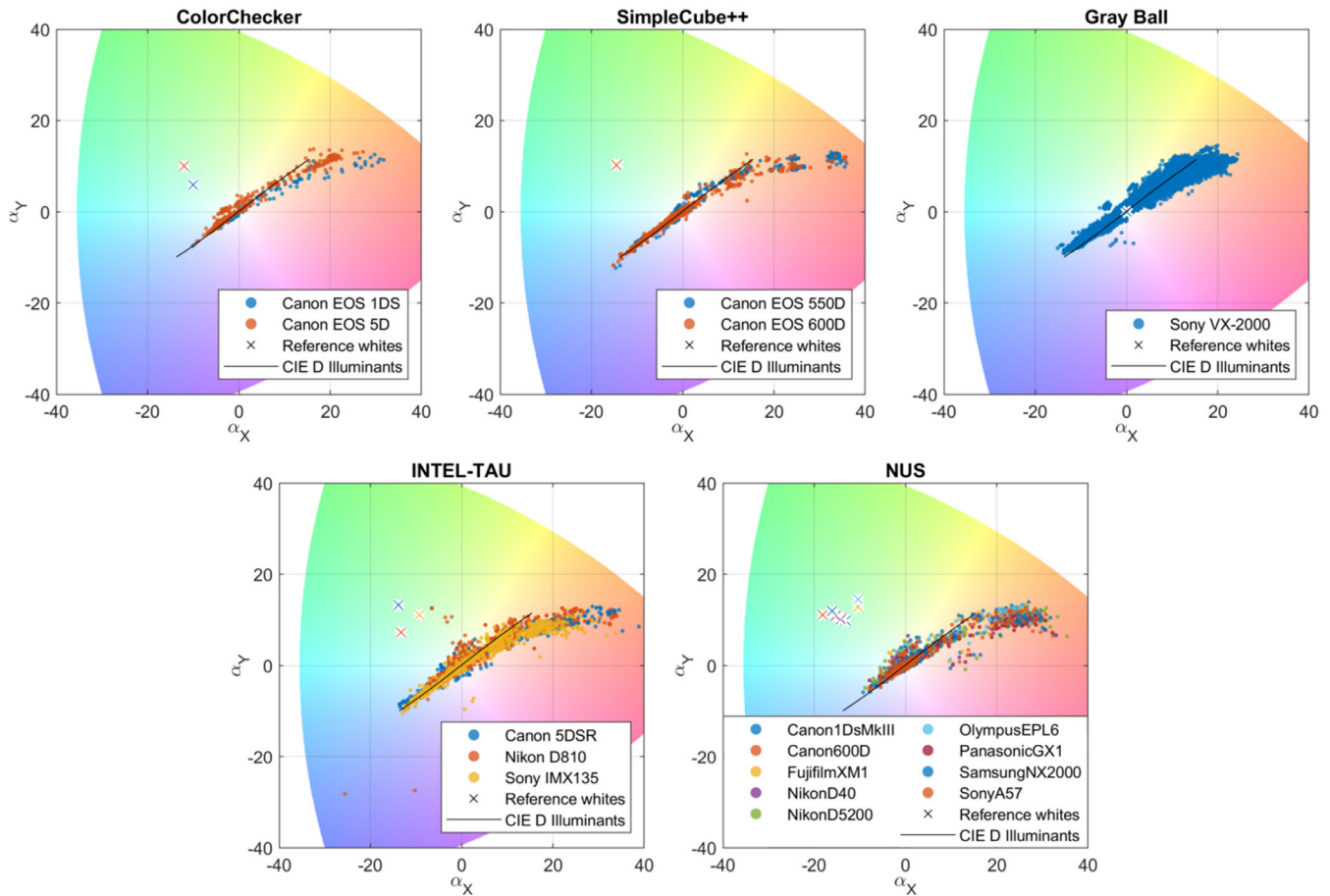


FIGURE 2 Illuminant distributions for five single-illuminant automatic white balance datasets, normalized by sensor-specific reference whites

reference white for each camera are documented in Table 3 in a normalized form such that the green channel G is equal to 1. Among the considered datasets, only INTEL-TAU provides a colorimetric characterization of the employed cameras. This can be exploited to accurately compute the corresponding reference white using the D55 spectrum from the Light Spectral Power Distribution Database.³⁷

In Figure 2, we plot the normalized datasets in the Angle-Retaining Chromaticity diagram³⁸ to avoid any representation-specific distortion of the data. The Cartesian coordinates α_X and α_Y are here used to define a frame of reference in a two-dimensional space where Euclidean distances are highly-correlated to angular distances from the original RAW RGB space. They are in turn obtained from polar coordinates α_A and α_R , which are described in Reference 38 as hue-like and saturation-like components. In the same plot, we also report the non-normalized reference whites are indicated by an \times symbol. For further reference, the CIE series D illuminants from D40 to D150 is also reported in each plot. We can observe that the illuminants in all analyzed datasets

roughly follow the distribution of daylight illuminants. All the datasets show additional data points at low correlated color temperatures (CCT). These are typically found in indoor scenarios illuminated by incandescent light sources. Among the five datasets, Gray Ball and INTEL-TAU exhibit the best coverage of high CCTs regions. This shows that these datasets contain several images with outdoor in-shadow surfaces. On a final note, with respect to the illuminants along the greenish to magenta-is directions, i.e. the direction orthogonal to the axis defined by CIE series D illuminants, we can see that they are poorly represented by all the considered datasets. We argue that collecting data for such non-natural light sources could be useful for investigating human perception on artificial lights.

3.1.2 | Multi-illuminant datasets

In terms of distribution of the involved chromaticities, the MIST dataset is characterized by a discrete set of individual illuminants, due to its synthetic nature, which are

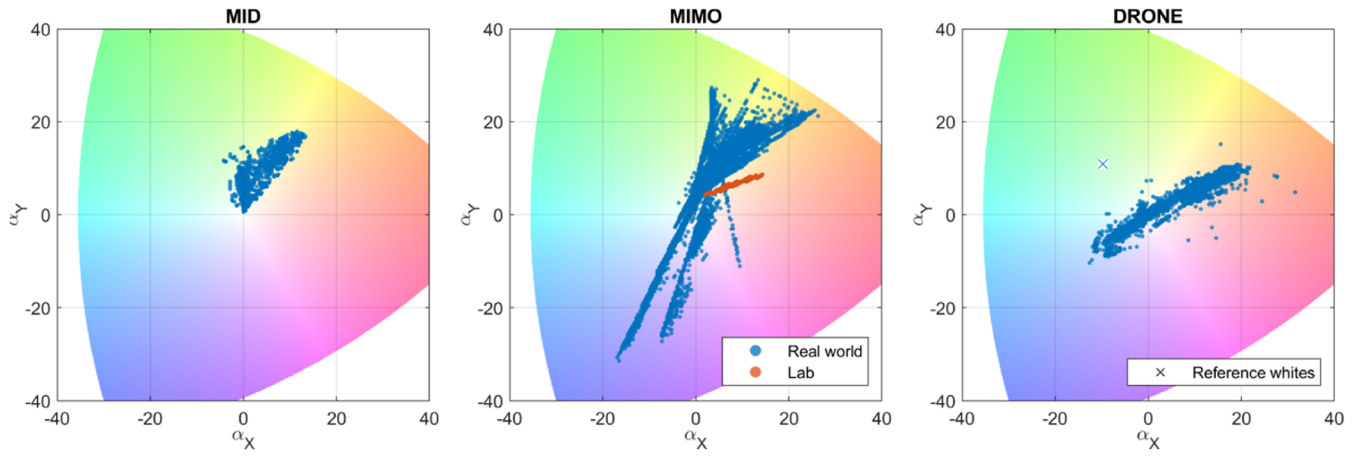


FIGURE 3 Illuminant distributions for three multi-illuminant automatic white balance datasets. Sensor-specific normalization was available and applied only to the DRONE dataset

TABLE 4 Summary statistics describing the distribution of single-illuminant and multi-illuminant datasets in terms of fitting bivariate Gaussian models and standard distance

Dataset	μ_{α_X}	μ_{α_Y}	Direction	σ_A	σ_B	STD
SimpleCube++	2.571	0.727	24.053°	11.565	1.188	11.623
ColorChecker	2.916	1.637	28.998°	11.174	1.116	11.220
Gray Ball	3.448	2.635	31.844°	9.977	1.107	10.038
INTEL-TAU	3.985	1.830	24.912°	9.975	1.447	10.078
NUS	4.598	2.253	23.073°	10.430	1.323	10.511
MID	4.151	9.530	47.990°	5.788	1.982	6.115
MIMO	4.208	5.642	59.479°	12.002	3.521	12.507
DRONE	6.943	3.377	29.515°	9.631	1.177	9.701
CIE D	-5.002	-3.645	36.717°	10.254	0.083	10.029

then blended into various combinations whenever multiple light sources illuminate the same surface. The underlying lights are standard illuminants D50, D65, A, F, 318 illuminants from the Illuminating Engineering Society,³⁹ as well as several equation-based illuminants with Correlated Color Temperature between 2000 and 10 000 K generated using the equations for CIE-D series and the black body radiator. Figure 3 offers a view of the illuminants in images from MID, MIMO, and DRONE, retrieved from the corresponding ground truth maps and tables. The procedure necessary for sensor-specific normalization previously presented on single-illuminant datasets was also employed here for the DRONE dataset alone (hand-picked D55-like sunlight reference corresponding to R/G = 0.5549 and B/G = 0.6489). Regarding MID and MIMO, no reliable information was available to unambiguously indicate a sensor-specific reference white for the involved cameras, thus preventing a direct comparison between the illuminant distributions between datasets. It is however possible to notice how, within the MIMO dataset, the illuminants of the “Real world” subset cover a considerably wider range of chromaticities than those of the “Lab” subset.

Summary statistics of the datasets illuminant distributions are presented in Table 4. By referring to chromatic coordinates in ARC diagram, we can analyze these distributions using metrics that imply data in a Euclidean space: as mentioned, this representation maps RGB angular distances (considered meaningful for AWB) into Euclidean distances. Following the example of an analysis performed on datasets for video AWB,³ we compute the standard distance⁴⁰ STD of the ARC datapoints, which is a bidimensional generalization of the standard deviation:

$$\text{STD} = \sqrt{\sum_{i=1}^N \frac{(\alpha_{X_i} - \overline{\alpha_X})^2}{N} + \sum_{i=1}^N \frac{(\alpha_{Y_i} - \overline{\alpha_Y})^2}{N}} \quad (2)$$

where $(\alpha_{X_i}, \alpha_{Y_i})$ are the ARC coordinates of the i th illuminant in a given dataset, and $(\overline{\alpha_X}, \overline{\alpha_Y})$ indicates the average of each coordinate. This metric provides a measure of how spread the illuminants for a given dataset are, taking into account density but neglecting cardinality, which is already accounted for in Tables 1 and 2.

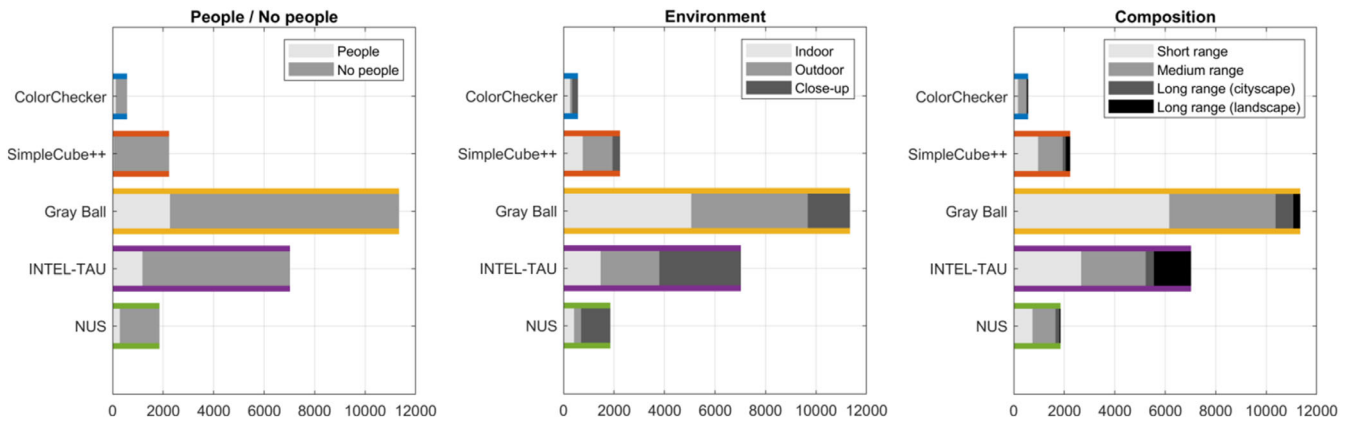


FIGURE 4 Image content distribution of the single-illuminant automatic white balance datasets, according to three different sets of classes

A more-detailed characterization can then be provided by fitting each distribution with a bivariate Gaussian model. In this case, we first process it via principal component analysis (PCA) in order to highlight the main axis of variation, described by the “Direction” column in Table 4, and subsequently extract two-dimensional moments μ and σ (average and standard deviation), where σ is related to the principal components A and B. These pieces of information can be put in comparison with the same analysis on the CIE D illuminants, in order to provide a frame of reference particularly for the μ and direction values.

3.2 | Image content

We have trained three convolutional neural networks on a proprietary annotated dataset in order to analyze the semantic content of the AWB images. The networks have been trained to classify images with respect to the presence of people, type of environment, and type of composition, respectively reaching 95.01%, 96.10%, and 97.50% accuracy on independent test sets. Using these networks, we then performed inference on white-balanced images from the five single-illuminant AWB datasets described in Section 2.1. We used this automatic annotation procedure to investigate the image contents of the AWB datasets. Figure 4 shows the aggregated results of the inference process of each network.

From the obtained results, we can see that the amount of images depicting human subjects is relatively small. This problem could be partially due to privacy-protecting instruments recently introduced in different countries such as the General Data Protection Regulation (also known as GDPR) in the European Union. It should be noted that some false positives may occur since the automatic annotation is not error-free. These false positives are related to statues,

posters, and other objects depicting people. The INTEL-TAU dataset depicts a relevant number of people; however, their faces are censored with an average-color mask. This could be a problem for some semantic-based AWB methods that explicitly rely on detecting human faces,¹⁵ and as such they cannot be trained or evaluated on this type of data. With respect to the depicted environment, that is, indoor, outdoor, or close-up, and the composition, that is, close-to-long range, all the datasets appear to be well balanced.

Contrarily to single-illuminant datasets, current multi-illuminant datasets are composed of relatively few scenes as reported in Table 2 (for a maximum of 31 scenes in the DRONE dataset³⁰). This allows for a direct inspection and analysis of the image content, instead of relying on aggregated statistics and automated analysis that may be subject to errors. None of the datasets depict any people. In terms of environment and composition, all four scenes of MID are strictly indoor of short range objects. The MIMO dataset depicts all short-range compositions. Of these, the 10 laboratory scenes are all indoor, while the real world scenes include three clearly outdoor setups, and the remaining 17 are shot indoor with occasional secondary sources from outside. Similarly, the DRONE dataset is composed of 10 clearly outdoor scenes, while the remaining 21 are indoor with an occasional mixture of indoor and outdoor light. For this dataset, the subjects of the photos are mostly in the medium range, with two short-range images. Finally, all three scenes of the MIST dataset depict medium range indoor setups.

3.3 | Shooting parameters and illumination levels

The next analysis refers to the illumination levels, that is, low-light to bright-light conditions. We characterized the

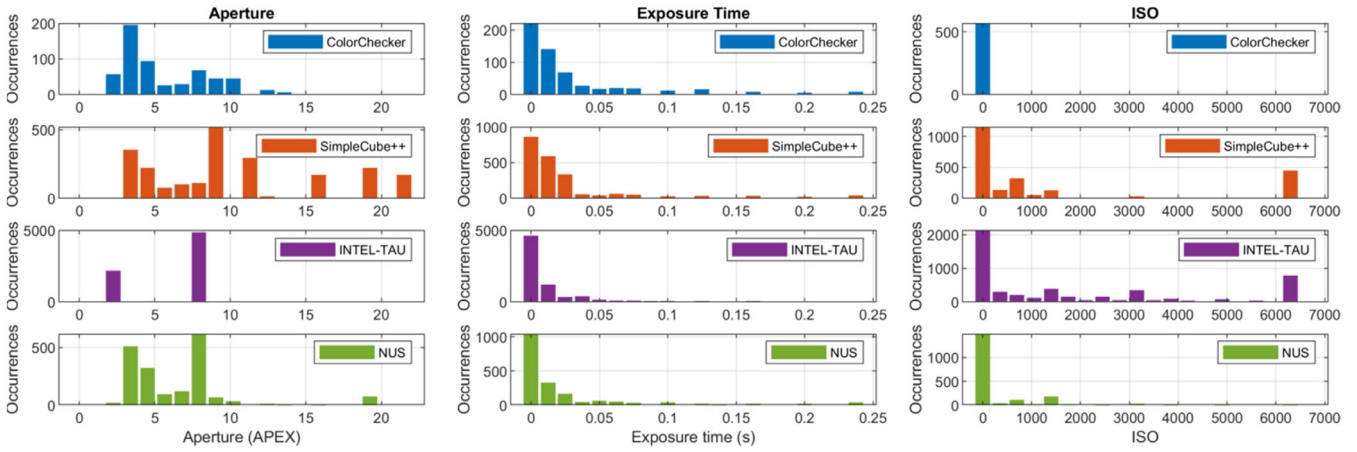


FIGURE 5 Shooting parameters distributions for the analyzed automatic white balance datasets: Aperture, Exposure Time, and ISO

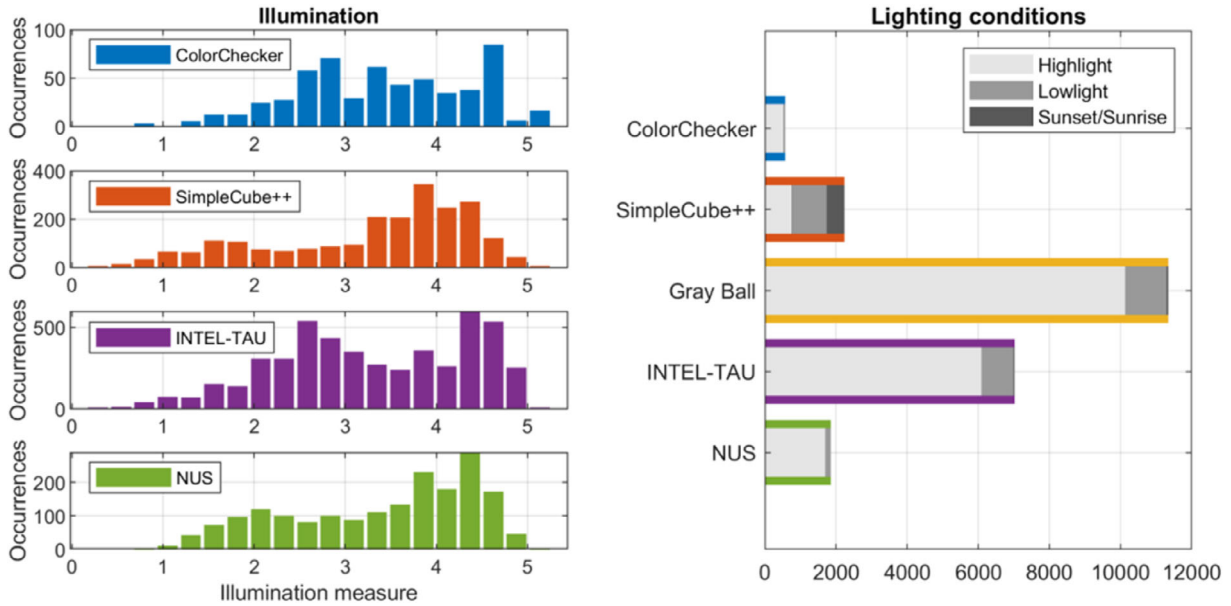


FIGURE 6 Illumination conditions computed by means of shooting parameters (left) and image analysis (right)

illumination levels of the five single-illuminant AWB datasets with two approaches.

The first approach is based on the EXIF data. We extracted the shooting parameters from the images in the dataset and analyzed their distribution, as reported in Figure 5. It must be noted that EXIF data are not available for the sRGB-encoded images in the Gray Ball dataset, so no analysis can be performed in this particular case. Moreover, the images captured with the Sony IMX132 camera in the INTEL-TAU dataset do not provide the ISO information, so the results are partial for this dataset. If we assume that the camera parameters were properly set during shooting, the aperture, exposure time, and ISO information can be used to approximate the scene illumination level. This can be achieved, for example, by using the following equation proposed by Le et al⁴¹:

$$I_{\text{measure}} = \log_{10} \left(\frac{\text{aperture}^2}{\text{exposure_time}} \right) + \log_{10} \left(\frac{250}{\text{ISO}} \right) \quad (3)$$

Figure 6 (left), shows the distribution of the I_{measure} for the four datasets with available EXIF data.

The second approach for illumination characterization is based on a classifier that is used to infer the scene illumination level in discrete terms. We trained a Convolutional Neural Network to classify an image into three illumination types: “Highlight,” “Lowlight,” and “Sunset/Sunrise.” The network was trained on a proprietary dataset of annotated images, reaching 98.33% accuracy on an independent test set. Prior to being classified, the images in the AWB datasets were sRGB-rendered and white-balanced in order to match the appearance expected by



FIGURE 7 Images from the SimpleCube++ dataset,²⁰ characterized by the corresponding lighting conditions and illumination measure (I_m)

TABLE 5 Cardinalities of single-illuminant datasets with respect to the class sets considered within this paper: people/no people, environment, composition, and lighting conditions

Set	Class	ColorChecker	SimpleCube++	Gray Ball	INTEL-TAU	NUS
People/no people	People	123	41	2269	1180	283
	No people	445	2193	9077	5842	1570
Environment	Indoor	254	758	5058	1466	415
	Outdoor	85	1182	4614	2327	279
	Close-up	229	294	1674	3229	1159
Composition	Short range	165	961	6157	2667	741
	Medium range	330	977	4215	2558	911
	Long range (cityscape)	32	121	701	325	158
	Long range (landscape)	41	175	273	1472	43
Lighting conditions	Highlight	516	743	10 131	6089	1695
	Lowlight	52	993	1165	911	158
	Sunset/sunrise	0	498	50	22	0

the CNN. The distributions of the inferred scene illumination are reported in Figure 6 (right).

For reference, Figure 7 shows example images from the SimpleCube++ dataset²⁰ identified in terms of lighting conditions and illumination measure.

It is possible to observe that both analyses show a predominance of bright light scenes. Middle-to-low light scenes are partially present in three of the five datasets: INTEL-TAU, Gray Ball and SimpleCube++. Also the ColorChecker dataset has some low-light scenes, but its overall low cardinality must also be taken into

consideration. It is worth noting that the Sunset/Sunrise class is scarcely represented in all datasets. Consumer devices equipped with AWB modules do not always produce optimal results in these kind of imaging conditions, therefore we argue that extreme imaging condition should be included in the AWB datasets, to better represent challenging real-life application cases.

The analyzed cardinalities in terms of image content (presence of people, environment, composition) and lighting conditions that are at the basis of Figures 4 and 6 (left) for single-illuminant datasets, are also documented

TABLE 6 Recovery and reproduction angular error statistics (average, median, and 99th percentile) for all combinations of considered single-illuminant AWB methods (rows) and datasets (columns)

Method	ColorChecker	SimpleCube++	Gray Ball	INTEL-TAU	NUS
Average recovery angular error					
WP	5.945	4.314	7.395	5.891	7.547
GW	4.716	3.185	6.814	4.937	4.224
SoG	4.178	3.117	6.273	4.752	5.069
GGW	4.364	3.131	6.892	5.103	4.686
GE1	3.890	3.059	6.101	4.599	3.913
GE2	4.070	3.172	5.717	4.194	3.771
FC4	1.922	2.116	5.523	2.571	3.587
GI	3.431	2.456	6.934	3.950	3.214
PCA	3.733	3.075	6.367	4.255	4.304
QU	3.289	2.387	6.463	3.474	3.418
SIIE	2.740	2.778	7.483	3.267	2.048
SURR	3.819	2.746	4.728	3.898	3.541
Average reproduction angular error					
WP	7.181	5.084	7.989	6.684	8.449
GW	5.775	4.128	7.203	6.166	5.469
SoG	5.285	3.885	6.823	5.789	6.205
GGW	5.410	3.950	7.417	6.259	5.921
GE1	4.874	3.909	6.583	5.711	5.048
GE2	5.136	4.083	6.242	5.264	4.843
FC4	2.506	2.723	6.106	3.318	4.526
GI	4.334	3.214	7.413	4.967	4.234
PCA	4.725	3.766	6.785	5.252	5.362
QU	4.148	3.110	6.971	4.361	4.468
SIIE	3.480	3.929	7.567	4.169	2.738
SURR	4.843	3.496	5.155	4.767	4.551
Median recovery angular error					
WP	3.839	1.869	5.837	3.567	4.577
GW	3.503	2.003	5.792	3.882	3.208
SoG	2.444	1.593	5.467	3.424	3.481
GGW	2.877	1.613	6.039	3.908	3.538
GE1	2.802	1.700	5.291	3.360	2.930
GE2	3.292	2.172	4.891	3.346	2.757
FC4	1.334	1.714	4.412	1.875	3.177
GI	2.235	1.256	5.197	2.364	2.134
PCA	2.323	1.487	5.624	2.882	2.952
QU	2.072	1.331	4.927	2.365	2.572
SIIE	1.939	1.472	6.599	2.340	1.503
SURR	2.403	1.378	3.853	2.406	2.351
Median reproduction angular error					
WP	4.877	2.501	6.020	4.351	5.782
GW	4.626	2.742	6.238	4.944	4.444

(Continues)

TABLE 6 (Continued)

Method	ColorChecker	SimpleCube++	Gray Ball	INTEL-TAU	NUS
SoG	3.224	2.176	5.885	4.262	4.562
GGW	3.529	2.268	6.527	4.898	4.729
GE1	3.587	2.318	5.663	4.213	3.943
GE2	4.158	2.905	5.238	4.228	3.572
FC4	1.622	2.133	4.602	2.359	3.670
GI	2.798	1.755	5.616	3.049	2.826
PCA	3.052	2.029	5.959	3.683	3.929
QU	2.571	1.835	5.165	3.024	3.488
SIIE	2.265	1.855	6.725	2.959	2.018
SURR	2.951	1.856	4.175	3.021	3.129
99th percentile recovery angular error					
WP	25.090	19.659	22.627	21.735	23.525
GW	17.343	13.633	22.106	18.089	15.807
SoG	18.780	17.248	19.870	18.602	18.557
GGW	18.913	16.698	21.865	19.174	16.904
GE1	15.306	16.460	19.798	18.279	14.891
GE2	14.424	14.642	17.489	15.455	16.202
FC4	8.434	9.764	19.623	11.772	11.631
GI	17.841	15.305	25.049	18.600	14.184
PCA	16.381	17.878	20.507	19.238	18.396
QU	14.602	13.575	21.709	15.789	14.495
SIIE	14.144	17.158	24.207	15.061	10.688
SURR	16.937	16.262	18.324	18.865	18.086
99th percentile angular error					
WP	29.533	23.692	25.432	26.606	29.806
GW	18.451	17.165	22.633	21.635	19.133
SoG	24.126	19.556	21.328	22.716	24.544
GGW	21.606	19.407	22.859	23.505	19.736
GE1	16.923	19.654	20.927	22.110	19.077
GE2	17.152	16.875	18.903	18.774	20.412
FC4	13.224	12.533	21.932	15.396	17.340
GI	21.765	18.469	25.294	21.604	18.528
PCA	22.128	19.751	21.797	22.440	20.408
QU	19.117	16.422	24.019	18.602	17.667
SIIE	18.612	24.574	24.281	19.225	14.561
SURR	23.094	19.333	19.591	21.690	20.503

in Table 5. We point out that these values are determined by automatic assessment: they are subject to error and should be considered as indicative of the real distributions. In addition to the presented considerations relative

to the strengths and shortcomings of each individual dataset, this overview could be used by researchers, and users in general, to inform the decision on which dataset to use when focusing on specific image categories.

4 | BIASES OF AWB METHODS FOR SINGLE ILLUMINANT

In this section, we select several methods for Automatic White Balance, and we analyze their performance in relationship to the content of the images. For this part of the analysis, we focus on single-illuminant AWB methods on single-illuminant AWB datasets.

4.1 | Analyzed methods

Our selection of AWB methods spans traditional solutions based on handcrafted features, as well as more recent approaches based on deep learning, in order to cover a variety of different approaches. All analyzed methods are sensor-independent. When training is necessary, we rely on official pre-trained models to ensure the best possible conditions.

In 2007, van de Weijer et al.⁴² proposed a framework to generalize multiple algorithms based on low-level image statistics. By modifying the free parameters within the framework, several known algorithms can be derived. Specifically, we considered six AWB algorithms by varying the parameters' configurations (Minkowski norm p and standard deviation σ) according to Reference 43:

Grey World (GW) : $p = 1, \sigma = 0$.

White Point (WP) : $p = \infty, \sigma = 0$.

Shades of Grey (SoG) : $p = 4, \sigma = 0$.

General Grey World (GGW) : $p = 9, \sigma = 9$.

1st order Grey Edge (GE1) : $p = 1, \sigma = 6$.

2nd order Grey Edge (GE2) : $p = 1, \sigma = 1$.

The Minkowski norm p could alternatively be computed according to a dynamic approach as suggested by Akbarinia et al.⁴⁴ The standard deviation parameter σ , defining the Gaussian filter applied by the underlying algorithms, is tightly related to the size of the input image. As such, we downscaled all images to have the maximum side be 256 pixels long for this family of AWB methods. An in-depth analysis of the relationship between p and σ parameters and methods performance is left for future developments.

Other approaches consider the entire color distribution. For example Cheng et al.²³ introduced an AWB algorithm based on PCA. They observed that the information



FIGURE 8 Example image from the ColorChecker dataset¹⁹ (left), with corresponding selection of achromatic pixel weights according to Quasi-Unsupervised Color Constancy method⁴⁷ (right)

provided by the PCA is comparable with more complex spatial analysis algorithms. The approach selects a percentage of dark and bright pixels for the computation. Here we select the percentage parameter to 3.5% following the best results obtained by the authors.

More recently, Qian et al,⁴⁵ proposed the Grayness Index (GI), a learning-free metric designed to identify neutral surfaces in an input image following the Dichromatic Reflection Model.⁴⁶ The authors showed that this index can be used to estimate a single illuminant as well as multiple illuminants. In this work, we used the default parameters from the implementation provided by the authors.

Convolutional neural networks (CNNs) have been successfully used in many application scenarios. Bianco et al developed a CNN-based Quasi-Unsupervised color constancy (QU) algorithm,⁴⁷ to detect achromatic pixels in color images. The network is trained without explicit AWB annotation. The only weak assumption is that the images are approximately balanced. The CNN model used in this work was trained on images from the ILSVRC2012 dataset of the ImageNet initiative.⁴⁸

Another approach exploiting convolutional networks is FC4 (Fully Convolutional Color Constancy with Confidence-weighted Pooling) by Hu et al.⁴⁹ An image is divided into patches and a neural network architecture assigns confidence weights to the patches according to the level on information they carry. The confidence values are then used to generate local estimates that are then merged into a global solution for the AWB problem. The official SqueezeNet-based implementation⁵⁰ is supplied with pre-trained models on each fold of the ColorChecker dataset,¹⁹ which we appropriately selected when

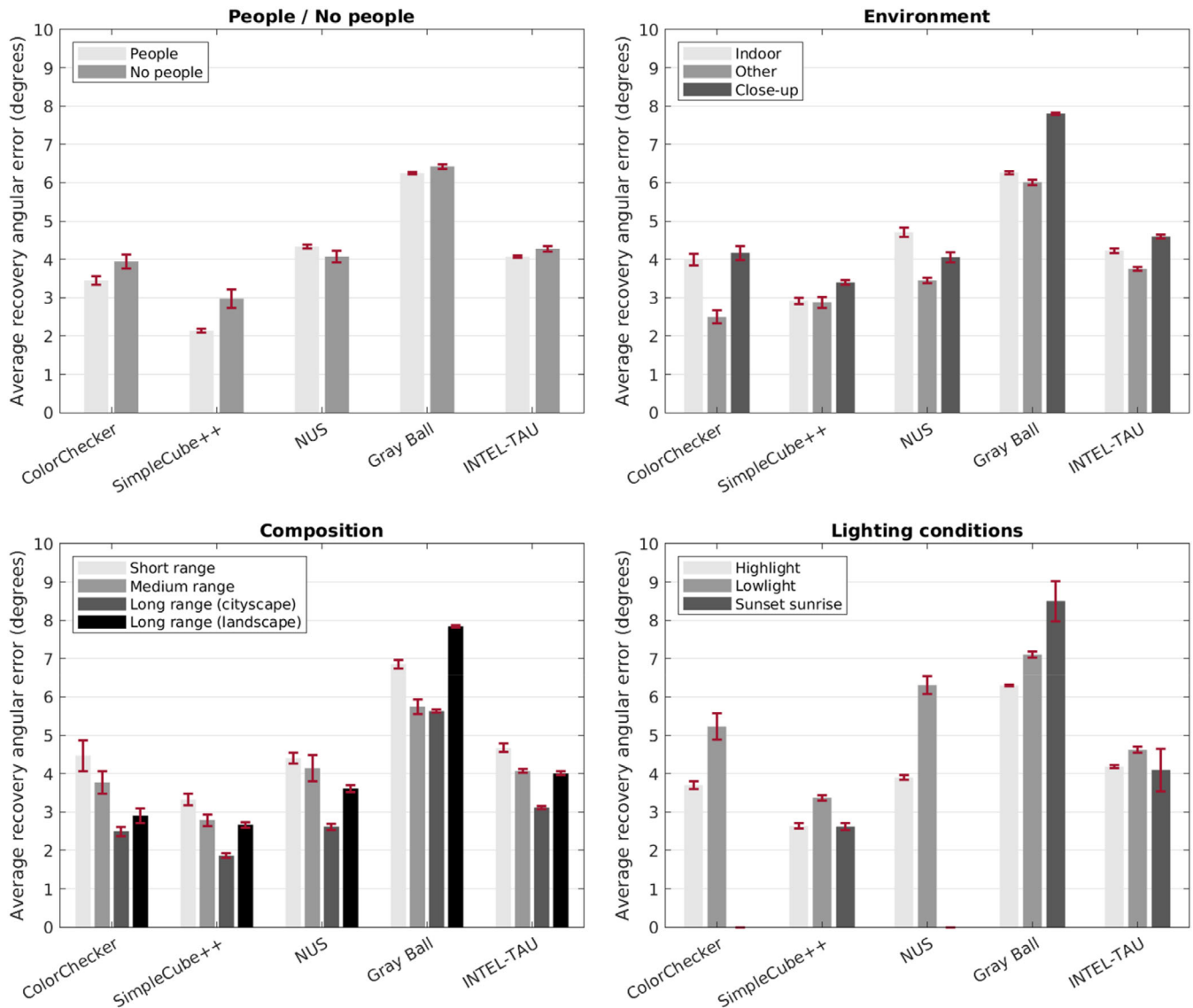


FIGURE 9 Average recovery angular error per class. Here are reported the average angular error of the considered method for each class of each label considered. From left to right, top to bottom are reported Composition label, Environment Label, Light conditions label and finally presence of people. Red bars represent the standard deviation of the errors done by the different methods, for each class on each dataset

testing on such datasets. For all other datasets, we used the model pretrained on “fold 2 and 0.”

Afifi et al² developed a learnable sensor-independent pseudo-RAW space to map the RGB values of any given camera, under the explicit assumption of input linear RAW-RGB images. The method is called Sensor-Independent Illumination Estimation (SIIE). We followed the method's guidelines and used the “MATLAB 2018b” model pretrained on the NUS²³ and Cube+⁵¹ datasets for testing on Gray Ball and INTEL-TAU datasets. We used the model pretrained on the NUS²³ and ColorChecker datasets for testing on the SimpleCube++. We finally relied on officially-reported estimations for testing on the ColorChecker and NUS datasets.

Akbarinia et al⁵² proposed Adaptive Surround Modulation (ASM): a solution inspired by the dynamic biological mechanism^{53,54} that correlates center-surround computations of local contrast with the receptive field size of visual neurons. Specifically, they use two overlapping asymmetric Gaussian kernels to model the visual neurons, and they weight their individual contributions according to the center-surround contrast. In this work, we used the default parameters from the implementation provided by the authors, and we used the publicly available estimations for the Gray Ball dataset.

Since several color constancy methods rely on the assumption of linear sensors, the following procedure³

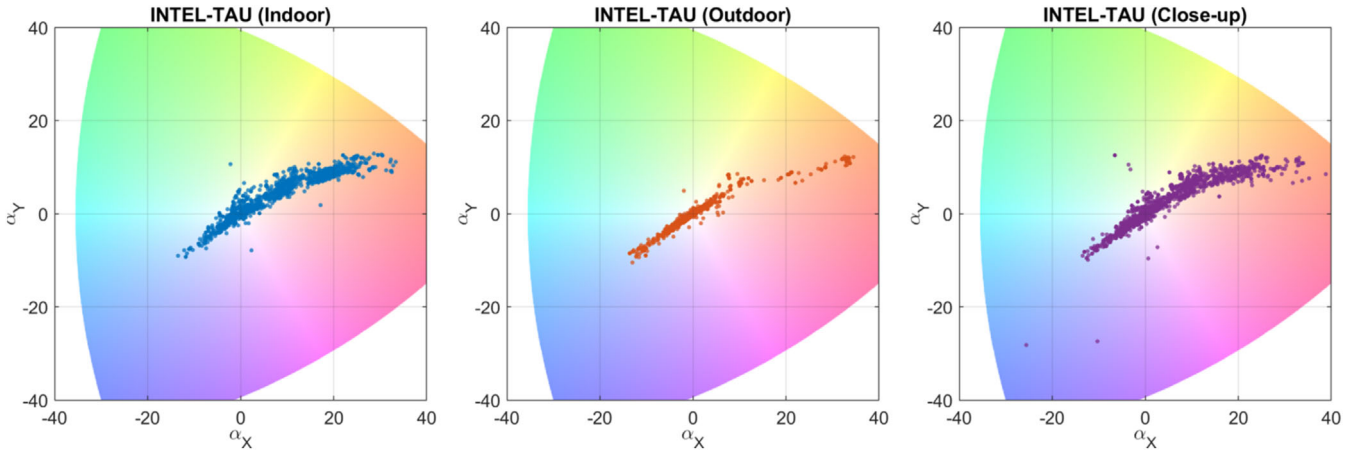


FIGURE 10 Illuminant chromaticity distribution of different “environment” cases for the INTEL-TAU dataset. Illuminants in the outdoor category are more compact than illuminants in the indoor category. The close-up category encompasses both types of scene

has been specifically applied to non-linear images from the Gray Ball dataset:

Linearize the image (gamma correction with $\gamma = 2.2$ ⁵⁵)

1. Estimate the illuminant
2. De-linearize the estimated illuminant ($\gamma = 1/2.2$)

As noted in Reference 3, this linearization strategy allows to process images that are closer to the RAW sensor data (although no guarantee on the actual gamma characterization is possible), while at the same time performing error analysis without modifying neither the AWB algorithms nor the official dataset ground truth.

4.2 | Performance per dataset

Error evaluation is here performed by resorting to the popular angle-based recovery angular error^{8,56} and reproduction angular error⁵⁷ between the estimated illuminant $V = (v_R, v_G, v_B)$ and the ground truth illuminant $U = (u_R, u_G, u_B)$:

$$\text{err}_{\text{rec}} = \arccos \left(\frac{\mathbf{U} \cdot \mathbf{V}}{|\mathbf{U}| |\mathbf{V}|} \right) = \arccos \left(\frac{\sum_i u_i v_i}{\sqrt{\sum_i u_i^2} \sqrt{\sum_i v_i^2}} \right) \quad (4)$$

$$\text{err}_{\text{rep}} = \arccos \left(\frac{u_{/V} \cdot (1, 1, 1)}{|u_{/V}| \sqrt{3}} \right) = \arccos \left(\frac{\sum_i u_i / v_i}{\sqrt{\sum_i u_i^2 / v_i^2} \sqrt{3}} \right) \quad (5)$$

The practical impact on human perception of variations in angular errors has been often debated. According to a survey by Gijzen et al,⁸ a deviation of 1° in angular error with the ground truth illuminant is considered below the threshold of what human vision commonly detects,⁹ while the $2^\circ \div 3^\circ$ range is considered perceivable but still acceptable.^{10,11} Hordley et al¹² identify a 2° recovery angular error as being acceptable in the context of complex scenes. These ranges should also be evaluated in the context of the data representation: when dealing with illuminant information in RGB form, variations in the RGB estimations are not linearly correlated with variations in angular errors, thus introducing a disconnect between representation and target. To this end, chromaticity diagrams such as the aforementioned Angle-Retaining Chromaticity can reduce this issue by formulating a data representation where Euclidean distances are highly correlated with RGB angular distances.

Table 6 presents a general overview of how the selected methods perform on the analyzed datasets for single-illuminant AWB. The average and median of the error distributions are reported, alongside the 99th percentile as a representative for the “worst-case” scenario. It is possible to observe that, in terms of datasets, the SimpleCube++ is on average the best handled across most analyzed methods, although such behavior is less noticeable when focusing on high error percentiles. Conversely, the hardest dataset appears to be Gray Ball according to both statistics. In this case, the non-RAW nature of the dataset is known to have a negative impact on many AWB methods.^{2,3} From the point of view of individual methods, FC4⁴⁹ reaches the best overall performance across most datasets, followed by either SIIE² or QU⁴⁷ depending on the statistics of interest.

TABLE 7 Average, median and 99th percentile of recovery angular error (degrees) over all datasets, for each method and identified class of images. Color coding is used separately for each row

Class	WP	GW	SoG	GCW	GE1	GE2	FC4	GI	PCA	QU	SIIE	ASM	Cardinality
Average recovery angular error													
People	9.001	4.541	5.268	5.600	4.758	4.851	4.825	4.457	5.130	6.164	5.340	3.960	3873
No people	6.122	5.859	5.373	5.777	5.198	4.806	3.892	5.368	5.187	4.570	5.174	4.210	19 033
Indoor	8.429	5.342	5.418	5.578	5.008	4.943	4.728	4.663	5.295	5.778	5.691	4.158	7929
Outdoor	4.605	5.687	5.182	5.777	4.947	4.498	3.235	6.130	5.045	3.877	5.025	4.094	8475
Other	7.001	5.929	5.504	5.914	5.493	5.067	4.286	4.692	5.206	4.949	4.838	4.277	6502
Short range	6.583	7.049	5.910	6.430	5.980	5.416	4.609	5.822	5.849	5.253	6.104	4.492	10 649
Medium range	7.434	4.312	4.950	5.091	4.408	4.476	3.742	4.275	4.740	4.807	4.470	3.908	8931
Long range (c.)	4.295	4.927	4.491	5.130	3.825	3.220	3.199	6.609	3.677	3.119	4.843	3.860	1323
Long range (l.)	4.596	4.500	4.787	5.447	4.619	4.166	3.014	5.243	4.542	3.923	3.915	3.807	2003
Highlight	6.543	5.764	5.403	5.858	5.169	4.840	4.029	5.387	5.252	4.812	5.257	4.220	19 067
Lowlight	7.401	5.308	5.441	5.525	5.178	4.877	4.462	4.628	5.029	5.393	5.257	4.056	3269
Sunset/Sunrise	4.253	3.251	3.277	3.302	3.294	3.542	2.386	2.771	3.516	2.597	3.047	3.085	570
Median recovery angular error													
People	7.947	3.698	4.115	4.547	3.722	3.920	3.704	3.128	3.852	4.207	3.782	2.852	3873
No people	3.841	4.747	4.218	4.663	4.078	3.840	2.752	3.625	3.969	3.127	4.205	3.011	19 033
Indoor	6.717	4.064	4.192	4.329	3.826	3.947	3.439	3.172	3.959	3.796	4.306	2.912	7929
Outdoor	2.737	4.863	4.151	4.900	4.049	3.718	2.361	4.325	4.033	2.888	4.692	3.100	8475
Other	5.181	4.758	4.299	4.643	4.229	3.963	3.086	3.116	3.836	3.469	3.189	2.892	6502
Short range	4.866	6.150	5.017	5.474	4.989	4.515	3.482	4.038	4.805	3.978	5.230	3.342	10 649
Medium range	4.733	3.582	3.683	4.005	3.374	3.568	2.521	3.040	3.404	2.930	3.361	2.826	8931
Long range (c.)	1.746	3.707	3.134	4.121	2.591	2.268	2.198	3.569	2.113	2.123	4.443	2.691	1323
Long range (l.)	2.844	3.271	3.501	4.400	3.544	3.582	2.093	3.451	3.188	2.715	2.679	2.495	2003
Highlight	4.439	4.671	4.261	4.764	4.067	3.879	2.871	3.691	4.048	3.311	4.352	3.082	19 067
Lowlight	4.636	4.127	4.286	4.381	4.126	3.955	3.173	2.974	3.758	3.575	3.554	2.716	3269
Sunset/Sunrise	1.907	1.351	1.363	1.412	1.365	2.277	1.824	1.084	1.399	1.219	1.430	1.427	570
99th percentile recovery angular error													
People	23.989	15.784	18.062	19.279	17.423	15.233	17.621	18.721	19.021	21.594	23.418	17.244	3873
No people	21.948	21.003	19.517	21.013	19.353	17.130	18.009	23.138	20.026	19.526	20.380	18.476	19 033
Indoor	23.617	21.906	18.551	19.742	17.984	16.652	17.793	20.534	19.759	21.410	25.764	18.362	7929

TABLE 7 (Continued)

Class	WP	GW	SoG	GGW	GE1	GE2	FC4	GI	PCA	QU	SIIE	ASM	Cardinality
Average recovery angular error													
Outdoor	19.060	19.066	19.211	20.857	18.221	15.862	16.105	24.557	19.894	17.645	16.562	17.788	8475
Other	22.055	20.286	19.714	21.279	20.209	17.540	18.688	20.975	20.044	19.055	20.212	19.176	6502
Short range	21.897	22.661	19.908	22.144	21.220	18.028	19.108	24.451	21.584	20.365	22.740	19.075	10 649
Medium range	23.210	14.882	18.013	18.269	16.349	15.319	16.888	18.841	18.751	20.562	20.825	17.962	8931
Long range (c.)	20.349	20.636	20.297	20.921	16.412	15.402	16.357	25.641	19.306	16.743	17.987	18.169	1323
Long range (l.)	19.089	16.998	20.012	21.515	16.293	13.475	14.812	24.047	19.575	16.714	16.083	17.600	2003
Highlight	22.163	20.672	19.158	20.728	18.834	16.686	17.994	23.084	19.939	20.156	21.322	18.334	19 067
Lowlight	23.730	19.739	19.736	19.257	18.320	17.268	17.622	19.827	19.184	20.156	21.189	18.133	3269
Sunset/sunrise	19.307	19.632	19.645	20.472	20.789	18.211	11.120	16.587	21.110	15.643	17.490	17.805	570

Any additional consideration should take into account the specific assumptions behind each method. The problem of Automatic White Balance, in fact, is known to be ill-posed and as such it cannot be analytically solved without additional assumptions on the imaged content. Among the methods based on the Edge-based color constancy,⁴² for example, the grey-world hypothesis assumes that the average reflectance in a scene is achromatic, the white patch hypothesis is based on the assumption that the maximum reflectance achieved for each of the color channels is equal, whereas the grey-edge hypothesis assumes that the average of the reflectance differences in a scene is achromatic. These simplistic low-level assumptions guarantee the efficiency of such methods, at the cost of generally lower performance: the grey-world hypothesis is easily broken when the scene is dominated by the intrinsic chromaticities of the surfaces, the white patch explicitly requires the presence of a bright neutral Lambertian surface in the scene, and the grey-edge can pick up misleading chromatic information from generally abrupt light intensity changes. The PCA method by Cheng et al²³ reformulates the assumption behind spatial domain methods, such as the grey-edge itself and Adaptive Surround Modulation,⁵² by postulating that most of the information coming from image derivatives can be directly extracted from the color distribution of the whole image, thus avoiding over-reliance on image gradients. The Greyness Index⁴⁵ and the Quasi-Unsupervised⁴⁷ methods are both designed to identify intrinsically achromatic pixels in the scene, whose RGB values are to be averaged for the purpose of illuminant estimation. As such, these methods are bound to work properly when such surfaces are indeed present. Nonetheless, the Quasi-Unsupervised method has been observed to exploit combinations of non-achromatic pixels to determine a global white point, as visualized in Figure 8, counteracting the intrinsic properties of the blue trash bin and the brown stones. More recent data-driven methods, such as FC4⁴⁹ and Sensor-Independent Illuminant Estimation,² implicitly rely on higher-level abstractions derived by the extraction of deep neural features. This allows them to formulate associations with the semantics of the depicted content, but also to exploit, and possibly over-rely upon, the aforementioned biases in the training data.

4.3 | Correlation between AWB performance and image content

For each of the considered AWB methods, we analyzed the recovery and reproduction angular error for different classes of image content identified in Section 3. The distributions are reported in Figure 9.

TABLE 8 Average, median and 99th percentile of reproduction angular error (degrees) over all datasets, for each method and identified class of images. Color coding is used separately for each row

Class	WP	GW	SoG	GGW	GE1	GE2	FC4	GI	PCA	QU	SIIE	ASM	Cardinality
Average reproduction angular error													
People	9.978	5.445	6.363	6.699	5.771	5.990	5.937	5.403	6.050	7.178	6.016	4.795	3873
No people	6.779	6.617	6.088	6.533	5.922	5.519	4.463	6.045	5.823	5.200	5.639	4.826	19 033
Indoor	9.382	6.243	6.386	6.536	5.889	5.896	5.708	5.492	6.090	6.648	6.299	4.866	7929
Outdoor	4.991	6.075	5.587	6.223	5.376	4.960	3.481	6.566	5.418	4.243	5.230	4.488	8475
Other	7.843	7.079	6.541	7.032	6.585	6.067	5.102	5.658	6.162	5.860	5.594	5.199	6502
Short range	7.282	7.890	6.651	7.214	6.735	6.125	5.252	6.545	6.515	5.899	6.528	5.150	10 649
Medium range	8.213	5.054	5.804	5.947	5.218	5.367	4.479	4.977	5.463	5.588	5.027	4.585	8931
Long range (c.)	4.694	5.323	4.869	5.549	4.172	3.645	3.499	6.952	3.979	3.431	5.131	4.170	1323
Long range (l.)	5.277	5.401	5.695	6.490	5.601	5.120	3.686	6.311	5.413	4.744	4.709	4.552	2003
Highlight	7.200	6.490	6.115	6.610	5.879	5.570	4.628	6.054	5.883	5.446	5.642	4.825	19 067
Lowlight	8.433	6.439	6.642	6.735	6.338	5.980	5.519	5.697	6.046	6.459	6.368	4.979	3269
Sunset/Sunrise	4.963	3.893	3.874	3.915	3.943	4.350	2.906	3.373	4.092	3.202	3.923	3.764	570
Median reproduction angular error													
People	8.696	4.407	4.936	5.459	4.506	4.809	4.476	3.847	4.571	5.130	4.523	3.509	3873
No people	4.344	5.567	4.806	5.378	4.774	4.511	3.131	4.193	4.572	3.620	4.731	3.499	19 033
Indoor	7.396	4.825	4.849	5.047	4.614	4.683	4.066	3.768	4.529	4.455	4.850	3.490	7929
Outdoor	2.988	5.383	4.529	5.352	4.455	4.236	2.614	4.805	4.461	3.163	5.033	3.440	8475
Other	6.196	5.989	5.310	5.795	5.331	4.869	3.751	3.916	4.767	4.336	4.006	3.652	6502
Short range	5.507	7.086	5.734	6.280	5.675	5.140	3.954	4.648	5.445	4.503	5.692	3.862	10 649
Medium range	5.370	4.166	4.214	4.598	3.940	4.152	2.898	3.657	3.972	3.461	3.933	3.310	8931
Long range (c.)	1.878	4.141	3.317	4.540	2.878	2.491	2.430	3.987	2.352	2.289	4.773	2.943	1323
Long range (l.)	3.505	3.985	4.299	5.309	4.415	4.471	2.535	4.513	4.068	3.395	3.411	3.058	2003
Highlight	4.957	5.467	4.849	5.460	4.747	4.565	3.247	4.263	4.642	3.791	4.823	3.556	19 067
Lowlight	5.533	5.030	5.235	5.489	5.115	4.789	3.890	3.844	4.705	4.478	4.427	3.445	3269
Sunset/Sunrise	2.429	1.807	1.736	1.781	1.828	2.992	2.219	1.515	1.878	1.588	1.843	1.928	570
99th percentile reproduction angular error													
People	9.978	5.445	6.363	6.699	5.771	5.990	5.937	5.403	6.050	7.178	6.016	4.795	3873
No people	6.779	6.617	6.088	6.533	5.922	5.519	4.463	6.045	5.823	5.200	5.639	4.826	19 033
Indoor	27.645	22.756	22.096	22.474	20.281	18.865	20.915	22.541	22.063	24.027	25.138	20.623	7929

TABLE 8 (Continued)

99th percentile reproduction angular error													
Outdoor	20.999	19.519	21.009	21.864	19.026	16.982	17.543	24.327	20.564	18.779	17.743	18.969	8475
Other	25.522	22.486	22.797	23.576	23.365	20.291	20.843	23.482	22.696	21.796	21.824	21.791	6502
Short range	24.917	23.746	22.363	23.898	23.034	20.043	21.379	25.190	23.445	23.051	24.167	21.399	10 649
Medium range	27.699	17.365	21.193	21.342	19.350	18.241	20.023	20.630	20.618	23.318	22.088	19.536	8931
Long range (c.)	22.271	21.486	22.405	21.773	18.699	17.213	18.064	25.375	20.648	18.182	18.719	19.629	1323
Long range (l.)	20.386	18.982	21.034	22.185	18.487	16.196	17.530	24.184	20.910	18.621	19.469	19.139	2003
Highlight	25.490	21.960	21.702	22.652	21.131	18.700	20.186	24.178	21.785	22.852	21.800	20.417	19 067
Lowlight	29.566	21.046	23.788	21.748	21.110	19.884	20.769	22.143	21.332	23.115	24.754	21.008	3269
Sunset/Sunrise	22.462	21.412	21.093	23.203	21.782	18.648	13.448	19.097	23.230	18.222	23.456	18.536	570

On average, the Gray Ball dataset is the one exhibiting the worst performance. As previously noted, this dataset is the only one not providing RAW file images, which may be the cause of such bad performance from all of the methods. Additionally, the images of the Gray Ball are of lower quality with respect to the other datasets, a property that may reduce the effectiveness of the automatic tagging procedure as well.

For what concerns the Composition set, all datasets show a consistent situation: we can observe that with images classified as long-range (both cityscapes and landscapes) the methods tend to perform better with respect to the medium range and short range images. This behavior, at least for the short range class, can be related to the fact that they have less context information, which would help in the analysis of the scene. A more-detailed analysis on image composition could potentially be performed by resorting to models for pixel-precise distance estimation,⁵⁸ providing information about different subjects depicted in a single image.

Regarding the Environment set, close-up images have less context for proper illuminant estimation, similarly to the short range class in the composition set, and as such generally worse performance can be observed. Additionally, the category of outdoor illuminants is in general relatively compact (especially assuming a daylight scenario), whereas indoor scenes can be illuminated by a wider variety of sources. This bias, experimentally observed on the distribution of the INTEL-TAU dataset in Figure 10, provides a possible explanation for the overall better performance of AWB methods on outdoor scenes.

Considering the light conditions of the images in the datasets, overall worse results can be observed when performing illuminant estimation on images identified as low-light. Here, the higher error can be related to two possible aspects. The first one can be the lack of ambient light: the absence of a global light source, such as the sun in overcast outdoor conditions, introduces the problem of having parts of the scene illuminated by different light sources with different chromaticity, thus breaking the implicit assumption of single-illuminant AWB methods. A second aspect which may cause the reduction in performance can be the lower numerical precision related to the dark pixels in the low light images.

Finally, for what concerns the presence of people in images, the performance on the different datasets are better when people are present in the scene. These results are in line with the assumption from Bianco et al,¹⁵ whom in their work explicitly exploited the presence of faces of people in images to estimate illuminant.

In Tables 7 and 8 are reported the average, median and 99th percentile of the recovery and reproduction error distributions respectively, of each method

considered for each identified class. The number of images per class is also reported in each table under the “Cardinality” column. From these tables it is possible to notice how two methods show the best performance on most of the identified classes: FC4 and ASM. It is also important to notice that the Gray Ball dataset have a much higher number of images with respect to the rest of the dataset: this fact explains the average better results of ASM and the 99th percentile performance of 2nd order Grey Edge, which show high performance on this dataset as reported in Table 6. In general, it is possible to notice the difference, in terms of performance, between traditional lower-level methods, and the more recent methods like SIIE, FC4 and Quasi-Unsupervised. It is also possible to confirm the same considerations that emerged by analyzing Figure 9.

5 | CONCLUSIONS

In this paper, we have compared the most popular datasets for Automatic White Balance in terms of illuminant distribution, image content, and shooting parameters. We have provided a detailed analysis and highlighted the individual shortcomings of each dataset. Deficiencies have been found across all datasets, such as a lack of images illuminated with artificial light sources and/or low-light images, indicating a direction for future data collection.

A selection of AWB methods from the state of the art has also been exploited to analyze the correlations between methods performance and image content. This analysis highlighted few classes for which the illuminant estimation procedure is harder. In particular, pictures taken in indoor scenarios or close-up shooting conditions appear to be harder than landscape or wider field of view ones. That happens also for pictures taken in low-light conditions with respect to the ones taken in highlight conditions. The presence of people in the images seems to help the illuminant estimation process; however, the number of images labeled with people presence for each of the considered dataset is low, so this conclusion cannot be considered definitive.

From these analysis, some suggestions to improve the performance of the algorithms trained on a single dataset can be made. For example, dividing the training and test images not in a random way but in such a way as to have in testing illuminations “never seen” in the training phase, or vice versa balancing them in training set so as not to give bias to the estimation method. Similar considerations can be made on the shooting parameters. We know, in fact, that semantic content is increasingly being used by methods of Automatic White Balance. This is a double-edged sword: on one hand, the algorithm can

leverage known information to better process the images and improve the final output, but on the other hand, it struggles in generalizing to different scenarios.

One possible way to tackle these problems is to consider the opportunity to merge multiple datasets into a more-complete set of images. For sensor-dependent AWB methods, this type of fusion can only be exploited by bringing the datasets into a common representation. To this extent, a set of reference white points has been compiled for each involved sensor, and shared within this manuscript. Also, merging different datasets can mitigate the bias in image content, providing a larger set of application scenarios to be addressed by the AWB algorithms.

The topic of multi-illuminant estimation in AWB is capturing a growing interest, with the emergence of appropriately annotated datasets as documented within this paper. The field is currently in its infancy, but we hypothesize a gradual shift toward more methods that produce dense illuminant estimation maps⁵⁹ and/or intrinsic decomposition,^{60,61} as covered in the past by methods inspired by the Retinex model.⁶² One of the main obstacles to multi-illuminant estimation and intrinsic decomposition lies in the collection of appropriately-annotated datasets. To this extent, the adoption of synthetic dataset generation^{63,64} provides a possible solution, as here anticipated in the case of the MIST dataset.²⁹

AUTHOR CONTRIBUTION

Marco Buzzelli: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Project Administration; Software; Supervision; Validation; Visualization; Writing – Original Draft Preparation; Writing – Review & Editing. **Simone Zini:** Data Curation; Formal Analysis; Software; Validation; Visualization; Writing – Original Draft Preparation; Writing – Review & Editing. **Simone Bianco:** Investigation; Project Administration; Resources; Writing – Review & Editing. **Gianluigi Ciocca:** Formal Analysis; Investigation; Project Administration; Resources; Writing – Review & Editing. **Raimondo Schettini:** Investigation; Project Administration; Resources; Supervision; Writing – Review & Editing. **Mikhail K. Tchobanou:** Investigation; Project Administration; Supervision; Writing – Review & Editing.

ACKNOWLEDGMENTS

This work is part of a research collaboration between the Imaging and Vision Laboratory from University of Milano - Bicocca, and Huawei Technologies Co. Ltd. Open Access Funding provided by Università degli Studi di Milano-Bicocca within the CRUI-CARE Agreement.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Marco Buzzelli  <https://orcid.org/0000-0003-1138-3345>
 Simone Zini  <https://orcid.org/0000-0002-8505-1581>
 Simone Bianco  <https://orcid.org/0000-0002-7070-1545>
 Gianluigi Ciocca  <https://orcid.org/0000-0003-2878-2131>
 Raimondo Schettini  <https://orcid.org/0000-0001-7461-1451>

REFERENCES

- [1] Foster DH. Does colour constancy exist? *Trends Cogn Sci*. 2003;7(10):439-443.
- [2] Afifi M, Brown MS. Sensor-independent illumination estimation for DNN models. In: *30th British Machine Vision Conference 2019 (BMVC 2019)*, p. 282. BMVA Press; 2019.
- [3] Buzzelli M, Erba I. On the evaluation of temporal and spatial stability of color constancy algorithms. *JOSA A*. 2021;38(9):1349-1356.
- [4] Hultgren BO, Hertel DW. Megapixel mythology and photospace: estimating photospace for camera phones from large image sets. In: SP Farnand and F Gaykema (Eds.), *Image Quality and System Performance V*. Vol. 6808, pp. 433-442. SPIE; 2008.
- [5] Hudelist MA, Schoeffmann K, Ahlströmy D, Lux M. How many, what and why? Visual media statistics on smartphones and tablets. In: *2015 IEEE International Conference on Multi-media & Expo Workshops (ICMEW)*. IEEE; 2015:1-6.
- [6] Ferwerda B, Tkalcic M. Predicting users' personality from instagram pictures: using visual and/or content features? In: *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. 2018:157-161.
- [7] GingerComms, HONOR 9X Research Reveals Europeans Now Post a Staggering 597 Photos of Themselves Every Year; 2019. Accessed on February 28, 2022, <https://www.hihonor.com/au/news/honor-9x-research-reveals-europeans-now-post-a-staggering597-photos-of-themselves-every-year/>
- [8] Gijssenij A, Gevers T, Lucassen MP. Perceptual analysis of distance measures for color constancy algorithms. *JOSA A*. 2009;26(10):2243-2256.
- [9] Funt B, Barnard K, Martin L. Is machine colour constancy good enough? In: *European Conference on Computer Vision* Springer; 1998:445-459.
- [10] Finlayson GD, Hordley SD, Morovic P. Colour constancy using the chromagenic constraint. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE; 2005:1079-1086.
- [11] Fredembach C, Finlayson G. Bright chromagenic algorithm for illuminant estimation. *J Imaging Sci Technol*. 2008;52(4):40906-40901.
- [12] Hordley SD. Scene illuminant estimation: past, present, and future. *Color Research & Application: endorsed by InterSociety color council, the colour group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of color, the Swedish colour Centre Foundation, colour Society of Australia. Centre Français de la Couleur*. 2006;31(4):303-314.
- [13] Bianco S, Ciocca G, Cusano C, Schettini R. Automatic color constancy algorithm selection and combination. *Pattern Recogn*. 2010;43(3):695-705.
- [14] Bianco S, Ciocca G, Cusano C, Schettini R. Improving color constancy using indoor-outdoor image classification. *IEEE Trans Image Process*. 2008;17(12):2381-2392.
- [15] Bianco S, Schettini R. Color constancy using faces. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2012:65-72.
- [16] Bianco S, Cusano C, Schettini R. Single and multiple illuminant estimation using convolutional neural networks. *IEEE Trans Image Process*. 2017;26(9):4347-4362.
- [17] Buzzelli M, van de Weijer J, Schettini R. Learning illuminant estimation from object recognition. In: *2018 25th IEEE International Conference on Image Processing (ICIP) IEEE*; 2018:3234-3238.
- [18] Bianco S, Buzzelli M, Ciocca G, Schettini R, Tchobanou M, Zini S. Analysis of biases in automatic white balance datasets. In: *Proceedings of the International Colour Association (AIC) Conference*; 2021:233-238.
- [19] Gehler PV, Rother C, Blake A, Minka T, Sharp T. Bayesian color constancy revisited. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE; 2008:1-8.
- [20] Ershov E, Savchik A, Semenkov I, et al. The cube++ illumination estimation dataset. *IEEE Access*. 2020;8:227511-227527.
- [21] Ciurea F, Funt B. A large image database for color constancy research. In: *Color and Imaging Conference No. 1, Society for Imaging Science and Technology*; 2003:160-164.
- [22] Laakom F, Raitoharju J, Nikkanen J, Iosifidis A, Gabbouj M. INTEL TAU: A color constancy dataset. *IEEE Access*. 2021;9:39560-39567.
- [23] Cheng D, Prasad DK, Brown MS. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*. 2014;31(5):1049-1058.
- [24] Shi L, Funt B. Re-processed Version of the Gehler Color Constancy Dataset of 568 Images; 2019. Accessed on February 28, 2022, https://www2.cs.sfu.ca/~colour/data/shi_gehler/
- [25] Hemrit G, Finlayson GD, Gijssenij A, et al. Rehabilitating the colorchecker dataset for illuminant estimation. *Color and Imaging Conference No. 26, Society for Imaging Science and Technology*; 2018:350-353.
- [26] Aytakin Ç, Nikkanen J, Gabbouj M. A data set for camera-independent color constancy. In *IEEE Transactions on Image Processing*, vol. 27, no. 2; 2018:530-544.
- [27] Bleier M, Riess C, Beigpour S, et al. Color constancy and non-uniform illumination: can existing algorithms work? In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE; 2011:774-781.
- [28] Beigpour S, Riess C, Van De Weijer J, Angelopoulou E. Multi-illuminant estimation with conditional random fields. *IEEE Trans Image Process*. 2013;23(1):83-96.
- [29] Hao X, Funt B. A multi-illuminant synthetic image test set. *Color Res Appl*. 2020;45(6):1055-1066.
- [30] Aghaei H, Funt B. A flying gray ball multi-illuminant image dataset for color research. *Color and Imaging Conference No. 28, Society for Imaging Science and Technology*; 2020:142-149.

- [31] Cheng D, Abdelhamed A, Price B, Cohen S, Brown MS. Two illuminant estimation and user correction preference. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016:469–477.
- [32] Dang-Nguyen DT, Pasquini C, Conotter V, Boato G. RAISE: A raw images dataset for digital image forensics. In: *Proceedings of the 6th ACM Multimedia Systems Conference*; 2015:219–224.
- [33] Erhov E, Savchik A, Semenkov I, et al., 2nd International Illumination Estimation Challenge; 2020. Accessed on February 28, 2022, https://chromaticity.iitp.ru/iec_2.html.
- [34] von Kries J. Theoretische studien über die umstimmung des sehorgans. *Festschrift der Albrecht-Ludwigs-Universität*. Vol 32. 1902;145-158.
- [35] Logvinenko AD, Funt B, Mirzaei H, Tokunaga R. Rethinking colour constancy. *PLoS One*. 2015;10(9):e0135029.
- [36] Mardaljevic J. In: Shamey R, ed. *Daylighting*. Springer Berlin Heidelberg; 2019:1-15. doi:10.1007/978-3-642-27851-8_258-2
- [37] Roby J, Aubé M, LSPDD | Light Spectral Power Distribution Database; 2015. <https://lspdd.org/app/en/lamps?lampuse=Standard%20Illuminant>, (Accessed on February 28, 2022).
- [38] Buzzelli M, Bianco S, Schettini R. ARC: angle-retaining chromaticity diagram for color constancy error analysis. *JOSA A*. 2020;37(11):1721-1730.
- [39] Luedtke W, Livingston J. IES method for evaluating light source color rendition (technical memorandum). *Illumin Eng Soc*. IES TM-30-15; 2018.
- [40] Burt JE, Barber GM, Rigby DL. *Elementary Statistics for Geographers*. Guilford Press; 2009.
- [41] Le QT, Ladret P, Nguyen HT, Caplier A. Large field/close-up image classification: from simple to very complex features. In: *International Conference on Computer Analysis of Images and Patterns Springer*; 2019:532–543.
- [42] Van De Weijer J, Gevers T, Gijssenij A. Edge-based color constancy. *IEEE Trans Image Process*. 2007;16(9):2207-2214.
- [43] Bianco S, Cusano C, Schettini R. Color constancy using CNNs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*; 2015:81–89.
- [44] Akbarinia A, Parraga CA. Colour Constancy: biologically-inspired contrast variant pooling mechanism. In: *BMVC 2017: 28th British Machine Vision Conference*; 2017.
- [45] Qian Y, Kamarainen JK, Nikkanen J, Matas J. On finding gray pixels. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2019:8062–8070.
- [46] Shafer SA. Using color to separate reflection components. *Color Res Appl*. 1985;10(4):210-218.
- [47] Bianco S, Cusano C. Quasi-unsupervised color constancy. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2019:12212–12221.
- [48] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Computer Vis*. 2015;115(3):211-252.
- [49] Hu Y, Wang B, Lin S. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017:4085–4094.
- [50] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv preprint arXiv:160207360*. 2016.
- [51] Banić N, Koščević K, Lončarić S. Unsupervised learning for color constancy. *arXiv preprint arXiv:171200436*. 2017.
- [52] Akbarinia A, Parraga CA. Colour constancy beyond the classical receptive field. *IEEE Trans Pattern Anal Mach Intell*. 2017;40(9):2081-2094.
- [53] Shushruth S, Ichida JM, Levitt JB, Angelucci A. Comparison of spatial summation properties of neurons in macaque V1 and V2. *J Neurophysiol*. 2009;102(4):2069-2083.
- [54] Angelucci A, Shushruth S. Beyond the classical receptive field: surround modulation in primary visual cortex. *New Vis Neurosci*. 2013;425-444.
- [55] Gijssenij A, Gevers T, Results per dataset (recovery error) | Color Constancy; 2011. Accessed on July 27, 2021, http://colorconstancy.com/evaluation/results-per-dataset/index.html#sfugreyball_linear.
- [56] Hordley SD, Finlayson GD. Reevaluation of color constancy algorithm performance. *JOSA A*. 2006;23(5):1008-1020.
- [57] Finlayson GD, Zakizadeh R. Reproduction angular error: an improved performance metric for illuminant estimation. *Perception*. 2014;310(1):1-26.
- [58] Bianco S, Buzzelli M, Schettini R. A unifying representation for pixel-precise distance estimation. *Multimed Tools Appl*. 2019;78(10):13767-13786.
- [59] Buzzelli M, Riva R, Bianco S, Schettini R. Consensus-driven illuminant estimation with GANs. In: *Thirteenth International Conference on Machine Vision*. SPIE; 2021:578–584.
- [60] Boss M, Braun R, Jampani V, Barron JT, Liu C, Lensch H. Nerd: Neural reflectance decomposition from image collections. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2021:12684–12694.
- [61] Baslamisli AS, Das P, Le HA, Karaoglu S, Gevers T. Shading-Net: image intrinsics by fine-grained shading decomposition. *Int J Computer Vis*. 2021;129(8):2445-2473.
- [62] Banić N, Lončarić S. Light random sprays Retinex: exploiting the noisy illumination estimation. *IEEE Signal Process Lett*. 2013;20(12):1240-1243.
- [63] Flachot A, Akbarinia A, Schütt HH, Fleming RW, Wichmann FA, Gegenfurtner KR. Deep neural models for color classification and color constancy. *J Vis*. 2022;22(4):17.
- [64] Gil Rodríguez R, Bayer F, Toscani M, Guarnera D, Guarnera GC, Gegenfurtner KR. Colour calibration of a head mounted display for colour vision research using virtual reality. *SN Comput Sci*. 2022;3(1):1-10.

AUTHOR BIOGRAPHIES



Marco Buzzelli obtained his Bachelor Degree and Master Degree in Computer Science respectively in 2012 and 2014, focusing on Image Processing and Computer Vision tasks. He received his PhD in Computer Science in 2019 at the University of Milano - Bicocca (Italy), where he is currently employed as a post-doctoral researcher. His main topics of research include characterization of digital

imaging devices, and object recognition in complex scenes.



Simone Zini obtained his Bachelor Degree and Master Degree in Computer Science at University of Milano - Bicocca (Italy), respectively in 2015 and 2018, focusing on image processing and enhancement. In 2018 he spent 5 months in Tokyo, Japan, for an internship period at NEC Corporation, working on Image denoising and enhancement, and obtained a research grant in the IVL lab of the University of Milano-Bicocca. He received his PhD in Computer Science in 2022. His main topics of research are machine learning, image processing and enhancement and computational photography.



Simone Bianco is associate professor of Computer Science at the University of Milano - Bicocca, holder of the Italian National Academic Qualification as Full Professor of Computer Engineering (09/H1) and Computer Science (01/B1). He is on Stanford University's World Ranking Scientists List for his achievements in Artificial Intelligence and Image Processing. His teaching and research interests include computer vision, artificial intelligence, machine learning, optimization algorithms applied in multimodal and multimedia applications. He is R&D Manager of the University of Milano - Bicocca spin off "Imaging and Vision Solutions," and member of ELLIS (European Laboratory for Learning and Intelligent Systems).



Gianluigi Ciocca obtained his Msc degree in Computer Science at the University of Milano in 1998. He has been a fellow at the Institute of Multimedia Information Technologies of the Italian National Research Council, where his research has focused on the development of systems for the management of

image and video databases and the development of new methodologies and algorithms for automatic indexing. Since 2003, he is at DISCo (Department of Informatics, Systems and Communication) of the University of Milano - Bicocca where he obtained his PhD in Computer Science in 2006. He is currently associate professor in Computer Science at DISCo. His current research fields are image and video analysis, pattern recognition, and classification. He is a member of NeuroMi: Milan Center for Neuroscience.



Raimondo Schettini is a professor at the University of Milano - Bicocca. He is Head of the Imaging and Vision Lab. He has been associated with the Italian National Research Council since 1987, where he led the color imaging lab from 1990 to 2002. He has been a team leader in several research projects and published more than 300 refereed papers and six patents about color reproduction, and image processing, analysis, and classification. He is a Fellow of the International Association of Pattern Recognition for his contributions to pattern recognition research and color image analysis.



Mikhail K. Tchobanou after an experience in academia as full professor at Moscow Power Engineering Institute (MPEI), joined Moscow Research Center Huawei in 2012. His research interests focus on image and color processing for embedded devices. He published more than 200 scientific papers and supervised several master and PhD students.

How to cite this article: Buzzelli M, Zini S, Bianco S, Ciocca G, Schettini R, Tchobanou MK. Analysis of biases in automatic white balance datasets and methods. *Color Res Appl.* 2022;1-23. doi:[10.1002/col.22822](https://doi.org/10.1002/col.22822)