Pattern Recognition 160 (2025) 111175

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Uncertainty estimation in color constancy

Marco Buzzelli [*,1], Simone Bianco [1]

*Department of Informatics Systems and Communication, University of Milano–Bicocca, Viale Sarca 336, Milan, 20126, Italy*

ARTICLE INFO

*Keywords:*
Uncertainty estimation
Color constancy
Automatic white balance
Illuminant estimation

ABSTRACT

Computational color constancy is an under-determined problem. As such, a key objective is to assign a level of uncertainty to the output illuminant estimations, which can significantly impact the reliability of the corrected images for downstream computer vision tasks. In this paper we present a formalization of uncertainty estimation in color constancy, and we define three forms of uncertainty that require at most one inference run to be estimated. The defined uncertainty estimators are applied to five different categories of color constancy algorithms. The experimental results on two standard datasets show a strong correlation between the estimated uncertainty and the illuminant estimation error. Furthermore, we show how color constancy algorithms can be cascaded leveraging the estimated uncertainty to provide more accurate illuminant estimates.

Computational color constancy refers to the problem of correcting the color cast of an image such that it appears as if it was acquired under a neutral or standard light source, called illuminant. A key challenge is the estimation of the existing illuminant from the image data alone, which is an under-determined problem [1] to the extent that human color constancy is known to fail too [2,3]. Therefore, by definition, all algorithms for computational color constancy operate in conditions of uncertainty, and must rely on additional assumptions on the input. Low level, statistics-based, algorithms make explicit assumptions about the statistical properties of natural scenes, and estimate the color of the illuminant as the deviation from such assumptions [4]. Most recent and effective algorithms are learning-based, and exploit models trained on handcrafted features extracted from the input image (e.g., [5–7]) or deep learning models (e.g., [8–10]). These methods operate more abstract levels of reasoning [11], and are expected to rely on assumptions based on the distribution of the training data [12,13].

The intrinsic ill-posed nature of the problem directly implies that information to answer the question "what is the illuminant?" is fundamentally missing, and we intend to convert this lack of information into a piece of information itself, answering the question "how uncertain is the illuminant?". It also implies that, by definition, all algorithms for computational color constancy operate in conditions of uncertainty, to which their effectiveness is rooted: Morovič et al. [14] argue in fact that there exist uncertainty and variation in a variety of color-related problems: in color representation, in quantities from which colorimetry is computed, and in perceptual evaluation, to the point that the representation of color information should always go beyond a single punctual piece of information. It appears therefore fundamental to associate a level of uncertainty to illuminant estimations, as a way to guide the interpretation of the results, and to aid the algorithms' explainability: Explainable Artificial Intelligence (XAI) [15] is concerned with providing the tools for human understanding of a model's response: identifying the different sources for uncertainty allows therefore to realize whether a certain output is grounded on previous knowledge, or how sensitive it is to inconspicuous fluctuations in the input representation. This, in turn, is desired as a way to provide human-interpretable feedback for the model's improvement. Estimating the uncertainty of a predicted illuminant can also have a direct impact on downstream computer vision tasks. Zini et al. [16] demonstrate that incorrect color constancy can significantly reduce recognition accuracy, highlighting the connection between these tasks. Therefore, we can envision a scenario where, if the estimated uncertainty for the illuminant is high, image recognition is performed on a grayscale version of the image to minimize the impact of unreliable color information, or under multiple color perturbations to achieve a consensus-based recognition.

In the field of uncertainty estimation, Der Kiureghian and Ditlevsen [17] initially formalized the concepts of aleatoric and epistemic uncertainty for engineering modeling: aleatoric uncertainty is an umbrella term that refers to the impact of uncontrollable random phenomena in a model's functioning, whereas epistemic uncertainty is explicitly tied to lack of sufficient knowledge and as such it is assumed to be potentially controlled with additional data. Kendal and Gal [18] further distinguished between homoscedastic aleatoric uncertainty, which describes the model's inherent noise as source of uncertainty, and heteroscedastic aleatoric uncertainty, which is dependent on the individual input. Many computer vision problems are nowadays often

* Corresponding author.
*E-mail addresses:* marco.buzzelli@unimib.it (M. Buzzelli), simone.bianco@unimib.it (S. Bianco).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.patcog.2024.111175
Received 24 November 2023; Received in revised form 10 October 2024; Accepted 6 November 2024
Available online 15 November 2024

addressed with the aid of deep learning models based on convolutional neural networks, due to their demonstrated effectiveness [19,20], although with the drawbacks of limited interpretability with respect to the usage of handcrafted techniques, and lack of calibrated probabilistic predictions [21]. A common solution [18] to the latter problem is the adoption of Bayesian Neural Networks (BNN), which replace a deterministic network's weight parameters with distributions over the parameters themselves. Hernandez-Lobato and Adams [21], however, note that BNNs are in practice incompatible with large datasets and network sizes, and propose a probabilistic backpropagation procedure to mitigate this problem. The results are found to be competitive, but not on par, with traditional backpropagation. Another limitation of BNNs is identified by Gal and Ghahramani [22] in the resulting constraints to computational complexity and test accuracy. They therefore proposed to exploit dropout layers as an approximation for Bayesian inference. This approach is shown to be effective, but it is limited to neural architectures that are designed and trained with such layers. Alternatively, several sources [23,24] propose the usage of deep ensembles to model uncertainty in inference, possibly in combination with dropout. In general, deep ensemble solutions represent a state of the art alternative to BNNs as noted by Ovadia et al. [25], but require a number of inference runs to produce reliable uncertainty estimations, thus creating efficiency issues for real-time deployment.

In this work, we address the problem of uncertainty estimation in computational color constancy, providing a set of solutions that generalize to different models (including, but not limited to, deep learning), and that require a limited amount of inference runs. Specifically, we present a formalization of uncertainty estimation in color constancy and we define three different forms of uncertainty, as depicted in Fig. 1. The first one (aleatoric) models uncertainty as the deviation of the illuminant estimate that is caused by a variation of the image content, reflecting the inherent noise during the image capture process; the second one (epistemic) models the uncertainty associated to the bias in the training data, with the most common illuminants being the less uncertain, helping to measure confidence in illuminant estimates from unseen scenarios; the third one (intrinsic) learns to model uncertainty from the image itself exploiting image properties that suggest lack of information, e.g. the close-up of a surface where it is impossible to disentangle the illuminant and the reflectance.

It is possible to envision an application of the estimated uncertainty to the performance of color constancy: to this extent, we conduct in Section 3.3 two preliminary experiments. In the first experiment we aim at improving the computational performance, resorting to computation-heavy color constancy algorithms only when the uncertainty of lightweight ones exceeds a given threshold. In the second experiment, we aim at improving the estimation performance. To this extent, we estimate the illuminant using a primary color constancy algorithm: if its estimated uncertainty is above a predefined threshold, we query a secondary color constancy algorithm, and replace its estimation if we measure a decrease in uncertainty.

There has been little to no published research in uncertainty estimation specifically for the task of computational color constancy. One example is the work by Zakizadeh et al. [26], where challenging images are identified as potentially more uncertain and as such should be processed in a different way. This approach can be in hindsight reinterpreted as an example of extremely specialized binary uncertainty estimation, nonetheless providing a practical use case application for it. A second example is the work by Hu et al. [10], which introduces the concept of a confidence-weighted pooling layer to weight the activations of the last neural feature map to produce the final illuminant estimate. Although this representation might contain information about the uncertainty of the overall estimation, no procedure is described to quantify this value. Similarly Bianco and Cusano [27] estimate the illuminant as a weighted sum of the input pixels, where the weights are the output of the network. Also in this case no procedure is described to associate an uncertainty value on the basis of the weight map.

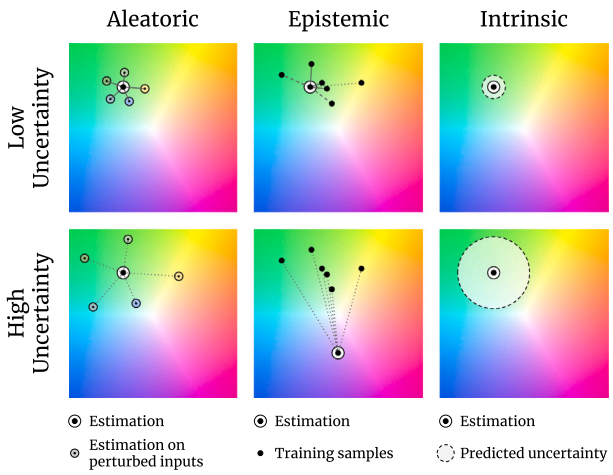


**Fig. 1.** Overview of the proposed methods for uncertainty estimation in color constancy. Aleatoric uncertainty measures the impact of input perturbations. Epistemic uncertainty quantifies the distance from a training set. Intrinsic uncertainty directly predicts an uncertainty radius.

Finally, Barron [28] casts the illuminant estimation problem as a 2-dimensional spatial localization task in a log-chrominance space, and assigns a likelihood score for all bins in the chroma histogram assuming that the highest-scoring bin is the color of the illuminant. Also in this case, the potential information contained in the score distribution is not exploited.

The main contributions of this work can be summarized as follows:

- we formalize the concept of uncertainty in color constancy, and define three forms of uncertainty that require at most one inference run in the trained model to be estimated;
- we show the applicability of our uncertainty estimators to different categories of color constancy methods, ranging from statistics-based approaches to deep CNN-based ones, and quantify their validity;
- we demonstrate an application of the usability of the estimated uncertainty to improve color constancy performance with a cascade approach.

## 1. Formalization and proposal

In this section we provide a formalization of uncertainty estimation in the context of computational color constancy, adapting the concepts of aleatoric and epistemic uncertainty to rely on a single inference run, and defining the concept of intrinsic uncertainty, which we present in the single shot and dual shot variants.

Let $x$ be an input color image with size $N \times M \times 3$ taking values in $\mathbb{R}^{N\times M\times 3}$. Let $f_{IE}(x;\theta)$ be an illuminant estimator $f_{IE} : \mathbb{R}^{N\times M\times 3} \rightarrow \mathbb{R}^3$, characterized by parameters $\theta$. Let $g_*(\cdot)$ be an uncertainty estimator that will be instantiated later with its input(s), domain, and co-domain. Let $y$ be the ground truth illuminant associated to the image $x$, and $\hat{y}$ the illuminant estimated by $f_{IE}$ for the image $x$, i.e. $\hat{y} = f_{IE}(x;\theta)$. Let $e$ be the recovery angular error between ground truth illuminant $y$ and estimated illuminant $\hat{y}$, i.e. $e = err(y,\hat{y})$.

The recovery angular error $err$ [29] quantifies the angular distance between two illuminants $U \in \mathbb{R}^3$ and $V \in \mathbb{R}^3$:

$$err(U,V) = \operatorname{acos}\left(\frac{U\cdot V}{\|U\|\,\|V\|}\right) = \operatorname{acos}\left(\frac{\sum_{i=1}^{3} u_i v_i}{\sqrt{\sum_{i=1}^{3} u_i^2}\sqrt{\sum_{i=1}^{3} v_i^2}}\right). \quad (1)$$

By definition, the angular error computation normalizes the input illuminants.

The accuracy of estimated uncertainty (UA) is here evaluated in terms of its correlation (*corr*) with the eventual recovery angular error for illuminant estimation, considering both Pearson and Spearman coefficients:

$$UA = corr\left(g_*(\cdot), err(y,\hat{y})\right). \tag{2}$$

The rationale is that, if an illuminant estimation is found to be particularly wrong (high error), it is desirable that it is associated to a high level of uncertainty. Additional metrics for the evaluation of uncertainty estimation will be defined in the experimental section.

### 1.1. Aleatoric uncertainty

Aleatoric uncertainty is here defined as the change in illuminant estimation resulting from a perturbation $\delta$ of the input, introduced by an operator $\oplus$:

$$g_A\left(f_{IE}(x;\theta), f_{IE}(x\oplus\delta;\theta)\right) : \mathbb{R}^3\cdot\mathbb{R}^3 \rightarrow \mathbb{R}. \tag{3}$$

The rationale is as follows: given a fixed perturbation of the input, a proportionally small change in the output suggests high confidence in the initial response, whereas a large change in the output suggests high uncertainty.

In our setup, we specifically consider *illuminant* perturbations, where $\oplus$ represents the introduction of an artificial illuminant $\delta = \{\delta_R, \delta_G, \delta_B\}$ to the input image. This is obtained by multiplying the green-normalized RGB illuminant through a von Kries-like [30] diagonal transform to each pixel $i$ of test image $x$:

$$\begin{aligned} x(i)\oplus\delta &= \left[x_R(i)\ x_G(i)\ x_B(i)\right] diag\left(\frac{\delta_R}{\delta_G}, 1, \frac{\delta_B}{\delta_G}\right) \\ &= \left[\frac{\delta_R}{\delta_G}x_R(i)\ x_G(i)\ \frac{\delta_B}{\delta_G}x_B(i)\right], \end{aligned} \tag{4}$$

i.e., each color channel is independently scaled with a factor that depends on the color of the illuminant.

The function $g_A$ used to measure the change in estimation is the recovery angular error, thus:

$$g_A := err\left(f_{IE}(x;\theta), f_{IE}(x\oplus\delta;\theta)\right). \tag{5}$$

Let $\Delta$ be a list of possible illuminant perturbations $\delta$: our goal is to find the perturbation which is the best predictor for uncertainty. We apply each perturbation $\delta\in\Delta$ to every test set image $x$ following Eq. (4). Then, for a given illuminant estimation method $f_{IE}$, we compute the set of illuminant estimation changes between $f_{IE}(x;\theta)$ and $f_{IE}(x\oplus\delta;\theta)$ $\forall\delta\in\Delta$, and optimize the best perturbation $\delta_0$:

$$\delta_0 = \text{argmax}_{\delta\in\Delta}\left(\text{UA}\left(e, err\left(f_{IE}(x;\theta), f_{IE}(x\oplus\delta;\theta)\right)\right)\right). \tag{6}$$

The discrete set of perturbations $\Delta$ considered for our experiments is presented in Section 2.1, and explored with grid-search via cross-validation.

### 1.2. Epistemic uncertainty

We define epistemic uncertainty as the degree to which the input image is dissimilar to a given knowledge base $K$, as described by feature extractors $h$ and $H$:

$$g_E\left(h(x), H(K)\right) : \mathbb{R}^3\cdot\mathbb{R}^{k\times 3} \rightarrow \mathbb{R}. \tag{7}$$

The rationale is as follows: an image that is similar to what has been observed during the training phase, suggests that the response of the illuminant estimator is grounded in knowledge and thus it can be associated to a high level of confidence. Conversely, an image that is particularly different from the training data, suggests high uncertainty in the produced illuminant estimation.

For the sake of generality, feature extractor $h$ could capture different aspects of the input image, from higher-level semantics to lower-level descriptors. In our setup, in practice, we use illuminant estimation itself as a feature extractor, therefore directly comparing images in terms of their illuminants. Accordingly, as knowledge base $K$ we consider the illuminant estimation training set having cardinality $|K| = k$, for which ground truth illuminants (as opposed to estimated) are available, taking the role of $H(K)$ in Eq. (7). We thus compare the illuminant estimation on the test image $x$, against all ground truth illuminants on the training set, to compute a distance vector $D$. For distance evaluation, we select the $p$th percentile ($ptile_p$) [31] of the resulting distribution, as a generalization of the minimum distance:

$$g_E := ptile_p\left(err\left(h(x), H(K)\right)\right), \tag{8}$$

Let $P$ be a list of possible percentiles. We select the $p$th percentile of distance vector $D$ as predictor for uncertainty. Determining the best percentile order $p_0$ is formalized as:

$$p_0 = \text{argmax}_{p\in P}\left(\text{UA}\left(e, ptile_p(D)\right)\right). \tag{9}$$

The discrete set of explored percentiles $P$ considered for our experiments is presented in Section 2.1, and explored with grid-search via cross-validation.

### 1.3. Intrinsic uncertainty

Intrinsic uncertainty is here defined as a property that can be estimated by direct analysis of the input image, assigning a degree of uncertainty to the illuminant estimated on the same image:

$$g_I(x;\theta_I). \tag{10}$$

This extraction relies on a machine learning model, trained with the objective of predicting the illuminant estimation error associated to an image, in line with Eq. (2):

$$g_I(x;\theta_I) = \hat{e} \approx e. \tag{11}$$

In the following, two variants of uncertainty extraction are introduced, called respectively single shot and dual shot.

#### 1.3.1. Single shot intrinsic uncertainty

In the single shot (SS) intrinsic uncertainty estimation, the original model for illuminant estimation is repurposed as a multi-task model $f_{IE}^{SS}$ for simultaneously estimating the scene illuminant and its uncertainty in a single shot:

$$g_I := f_{IE}^{SS} : \mathbb{R}^{N\times M\times 3} \rightarrow \mathbb{R}^4 = \left\{\mathbb{R}^3, \mathbb{R}\right\}, \tag{12}$$

so that $f_{IE}^{SS}(x;\theta_I) = \{\hat{y}, \hat{e}\}$. Multi-task learning has been shown to improve the generalization of the model by obtaining knowledge in related tasks that can serve as a further regularization [32]. In particular, we argue that uncertainty estimation and illuminant estimation are related tasks, and the first one can greatly benefit of the knowledge gained from the second one. As a multi-task learning problem, its training requires a total loss composed by three different terms. The first term $\mathcal{L}_{IE}$ is related to illuminant estimation:

$$\mathcal{L}_{IE} = err(y,\hat{y}) \tag{13}$$

and measures for each image in the batch the recovery angular error defined in Eq. (1) between the ground truth and the estimated illuminant. The second term is related to uncertainty estimation:

$$\mathcal{L}_{UE\text{-}L1} = d_1(e,\hat{e}) \tag{14}$$

and measures for each image in the batch the $L_1$ distance between the illuminant estimation error and the predicted one. The third term is also related to uncertainty estimation:

$$\mathcal{L}_{UE\text{-}C} = -|PCC(\hat{e}, e)| \tag{15}$$

and measures for each batch the absolute value of the Pearson correlation coefficient between the illuminant estimation errors and the predicted ones, following the provided definition of uncertainty accuracy (UA). The total loss $\mathcal{L}_{TOT}$ can then be written as:

$$\mathcal{L}_{TOT} = \lambda_{IE} \cdot \mathcal{L}_{IE} + \lambda_{UE\text{-}L1} \cdot \mathcal{L}_{UE\text{-}L1} + \lambda_{UE\text{-}C} \cdot \mathcal{L}_{UE\text{-}C}, \tag{16}$$

where $\lambda_{IE}$, $\lambda_{UE\text{-}L1}$, and $\lambda_{UE\text{-}C}$ are three scalars regulating the relative contributions of the respective terms in the loss.

#### 1.3.2. Dual shot intrinsic uncertainty

In the dual shot (DS) uncertainty estimation, the model $g_I$ is trained to predict the illuminant estimation error $e = err(y, \hat{y})$ resulting from the estimation $\hat{y} = f_{IE}(x;\theta)$. The model $g_I$ is the original model for illuminant estimation, repurposed for uncertainty estimation:

$$g_I := f_{IE}^{DS} : \mathbb{R}^{N\times M\times 3} \rightarrow \mathbb{R}, \tag{17}$$

so that $f_{IE}^{DS}(x;\theta_2) = \hat{e}$.

After we have trained the illuminant estimator $f_{IE}(x;\theta)$, that for each image $x_i$ in the dataset $\mathcal{X}$ produces an estimate $\hat{y}_i = f_{IE}(x_i;\theta)$, we train a second version of the estimator that is now repurposed for uncertainty estimation so that $f_{IE}^{DS}(x_i;\theta_2) = \hat{e}_i = err(y, \hat{y}_i)$. This dual shot estimator is trained using the total loss:

$$\mathcal{L}_{TOT} = \lambda_{UE\text{-}L1} \cdot \mathcal{L}_{UE\text{-}L1} + \lambda_{UE\text{-}C} \cdot \mathcal{L}_{UE\text{-}C}, \tag{18}$$

where $\mathcal{L}_{UE\text{-}L1}$ and $\mathcal{L}_{UE\text{-}C}$ are respectively defined in Eqs. (14) and (15). DS uncertainty estimation has the advantage over SS in not having to find a trade-off between illuminant and uncertainty estimation accuracy, since there are two sets of weights trained for the two estimation tasks. It also serves to prove if uncertainty estimation and illuminant estimation are related tasks, and to measure the performance improvement in the first task due to the knowledge coming from the second one.

### 1.4. Methods categorization

The proposed methods to estimate uncertainty in the domain of color constancy are characterized by significantly different approaches, which lead to different types of dependence from training bias and domain shift. Aleatoric uncertainty operates directly and solely on the input image, without the need for a training set. As such, it is not subject to influences related to domain shift. Epistemic uncertainty is explicitly designed to leverage the bias of a training dataset. As a consequence, any effect of domain shift is intrinsically tied to the method itself. Intrinsic uncertainty is potentially subject to suffering from an over dependence on training conditions. Therefore, it is more sensitive to domain shift effects.

## 2. Experimental setup

**Datasets** We use two real-world color constancy datasets to evaluate the performance of the different uncertainty estimation methods:

- The Gehler–Shi dataset [33] (also known as the ColorChecker dataset) with REC groundtruth [34] depicts a number of indoor and outdoor scenarios, and includes several human subjects in the shots.
- The Cube++ datasetdataset [35] includes ten-fold the number of images of Gehler–Shi, and also depicts a variety of scenes from indoor and outdoor setups.

For both datasets we masked all the color checker targets by setting the pixel values to (0, 0, 0) in RGB space.

**Illuminant estimation methods** Illuminant estimation algorithms in the literature can be grouped in different categories, on the basis of their assumptions and the techniques used for estimation. As stated in the previous sections, aleatoric uncertainty and epistemic uncertainty can be applied to any illuminant estimation algorithm without any modification. Intrinsic uncertainty estimation instead requires that the illuminant estimation algorithm is based on machine learning. In order to test and compare the different uncertainty estimation methods proposed in this paper, we select one illuminant estimation method for each of the most commonly used machine learning-based categories:

- Tree-based: [7]
- Deep CNN-based: FC$^4$ (SqueezeNet) [10]
- Shallow CNN-based: Convolutional Mean [36]
- Statistics-based: Corrected Moments [37]
- Combination-based (linear) : LMS committee [38]

The method by [7] is trained on full size images using the original implementation from the authors.

FC$^4$ [10] is trained on $512 \times 512$ images, using Adam as optimizer with a learning rate of 3e−4, a weight decay equal to 5e−5, for a total of 1000 epochs, with a batch size of 16. During training the images are augmented with a random rotation in $[-30°, 30°]$, with random crops having a scale factor in the range $[0.1, 1.00]$ and an aspect ratio in the range $[0.9, 1.1]$, and random horizontal flip with 0.5 probability. The unaugmented training set is used as validation data to select the best model. FC$^4$+SS is trained using the same hyperparameters of FC$^4$ with the difference that, after 1000 epochs of training the whole model, 500 additional epochs are spent to fine-tune only the final layer, mapping to the estimated uncertainty. FC$^4$+DS is trained in the exact same configuration of FC$^4$.

Convolutional Mean (Conv.Mean) [36] is trained with the same procedure for FC$^4$, but on $200 \times 200$ input images using a 5e−3 learning rate. Conv.Mean+SS and Conv.Mean+DS are respectively trained like FC$^4$+SS and FC$^4$+DS.

The Corrected Moments [37] variant implemented in this paper is the 9 Color-Edge moments, and the correction matrix is found by following the alternating least squares solution strategy proposed by the authors.

LMS Committee [38] combines six statistics-based algorithms that are instantiations of the Gray Edge framework [4]: Shades of Gray (SoG), General Gray World (gGW), Gray Edge 1st order (GE1), Gray Edge 2nd order (GE2), Gray World (GW), and White point (WP). The parameters of each algorithm are set as in [39], and the input images are resized with the longest side to 256. Before combining, each individual illuminant estimate is normalized to unitary norm, and the combination matrix is estimated using the standard pseudo-inverse.

### 2.1. Uncertainty implementation details

**Monte Carlo dropout** [22]. We implement Monte Carlo (MC) dropout uncertainty estimation as representative for state of the art deep ensembles. We perform $T = 100$ stochastic forward passes through the trained model, averaging the per-channel standard deviations of the illuminant estimations. Performance at different values of $T$ is reported in Appendix A.3.

**Aleatoric uncertainty**. We generate 629 illuminants, sampled uniformly in Angle-Retaining Chromaticity (ARC) [40]. Resorting to ARC representation allows us to cover the whole chromaticity diagram without biases towards specific regions. Each sampled ARC illuminant $\{\delta_{\alpha_A}, \delta_{\alpha_R}\}$ is then converted into a corresponding RGB illuminant $\{\delta_R, \delta_G, \delta_B\}$ before application via diagonal transform. In order to ensure a fair selection of the best perturbation $\delta$, cross-validation is used for the Gehler–Shi dataset (using the three official folds) and for the SimpleCube++ test set (using two folds). A visualization of the landscape of possible perturbations is provided in Appendix A.1.

**Epistemic uncertainty**. We evaluate the Euclidean distance between the ARC-encoded estimated illuminant $f_{IE}(x;\theta)$ and all ARC-encoded ground truth training illuminants $GT(K)$, producing a distance vector $D$:

$$D = dist_{Euc}\left(ARC\left(f_{IE}(x;\theta)\right), ARC\left(GT(K)\right)\right). \tag{19}$$

**Table 1**
Main characteristics of the considered uncertainty types.

| Uncertainty type | No. inference runs (IE+Uncert.) | Works on any IE model | Needs to retrain the IE model |
|---|---|---|---|
| MC dropout [22] | 1+T | No (dropout req.) | No |
| Aleatoric | 1+1 | Yes | No |
| Epistemic | 1 | Yes | No |
| Intrinsic | 1 (SS), 1+1 (DS) | No (SS), Yes (DS) | Yes |

The usage of the ARC representation allows us to efficiently compute at inference time a distance matrix based on Euclidean distances, as a proxy for angular distances in RGB space, which are commonly used in the comparison of illuminants. We consider $P = [0, 50]$ as the list of percentiles evaluated for best uncertainty predictor, whose behavior is explored in Appendix A.1. Cross-validation is also used to ensure a fair evaluation.

**Single shot intrinsic uncertainty**. The weights in the total loss defined in Eq. (16) are set as follows: $\lambda_{IE} = 1$, while the weights $\lambda_{UE\text{-}L1}$ and $\lambda_{UE\text{-}C}$ are set by performing multiobjective hyperparameter optimization on the validation set: we target performance on FC$^4$, selecting the Pareto-optimal configuration that produces the highest correlation, while limiting the angular error deterioration to 10% from the baseline model. The resulting optimized weights are $\lambda_{UE\text{-}L1} = 2.56e-5$ and $\lambda_{UE\text{-}C} = 2.55$.

**Dual shot intrinsic uncertainty** $\lambda_{UE\text{-}L1}$ and $\lambda_{UE\text{-}C}$ are set to the same values of the Single shot case.

The main characteristics of the uncertainty estimators considered in this work are reported in Table 1 in terms of: inference runs needed for illuminant and uncertainty estimation, compatibility with any illuminant estimation algorithm, and necessity to retrain a repurposed version of the illuminant estimation algorithm. The compatibility constraint of MC dropout is using a deep learning model designed and trained with a dropout layer. The same constraint also applies to more recent variants, e.g. MC DropBlock [41]. The constraint of Intrinsic (SS) is using the custom loss in Eq. (16) which, in our setup, excludes the algorithm in [7] since it builds one model for each of the illuminant components, and excludes the LMS Committee [38] since it would require replacing the LMS solver with a nonlinear optimizer.

## 3. Experimental results

The experimental results are reported in Table 2 in terms of angular error statistics (the lower the better) and correlation (the higher the better) including both Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient (SRCC). All statistics are averaged over three independent runs. Concerning the angular error we report the mean and median values as measures of the central tendency, the trimean which combines the median's emphasis on center values with the midhinge's attention to the extremes [42], and the 95th and 99th percentiles as statistics of the worst case performance which can have catastrophic effects on recognition performance if the images are used for downstream computer vision tasks, or human judgment if the images are for personal collection. The correlation coefficients are used to evaluate the quality of the estimated uncertainty instead of distance metric, since uncertainty can be later calibrated to assume values in the desired range. Correlations measure to which extent the estimated uncertainty can be calibrated with a linear function (PCC) or with a non-linear function (SRCC).

From the experimental results on the Gehler–Shi dataset [33] with REC groundtruth [34] reported in Table 2 we can notice that the method reaching the lowest angular error on the Mean, Median, and Trimean statistic is FC$^4$, while the lowest 95th and 99th percentile errors are reached by FC$^4$+SS, i.e. FC$^4$ with Single shot intrinsic uncertainty. Note that we report high-order percentiles as a generalization of the less-robust maximum error.

Recall that Dual shot trains the same model in two phases, generating two separate parameter sets for two different tasks: illuminant estimation and uncertainty estimation, whereas Single shot attempts to perform multitasking, learning a single set of features for both tasks. The fact that SS outperforms DS suggests that the two tasks are indeed related and benefit from shared knowledge. If DS had performed better, it would have indicated that the two problems are less interconnected. We can also observe how the highest correlations are obtained by MC dropout [22], epistemic uncertainty, or by SS intrinsic uncertainty. Furthermore, we can observe how in certain cases, applying MC dropout [22] or Epistemic uncertainty on a model trained with SS Intrinsic uncertainty is able to further improve the correlation. This can be observed for example for Corrected Moments [37] in terms of PCC, and Convolutional Mean [36] in terms of SRCC. These findings are also confirmed on the Simple Cube++ dataset [35], as reported in Table 3. Overall, epistemic uncertainty provides the best, or second best, results per correlation type in most configurations, with an average PCC of 0.4682 and SRCC of 0.5137, suggesting that it is a solid baseline choice for general situations. Additionally, the computation of epistemic uncertainty over models trained for Single shot uncertainty introduces an average improvement of 0.1787 points in PCC and 0.1274 points in SRCC.

In general we can observe that on the Simple Cube++ dataset the angular error is lower and the correlations are higher with respect to the Gehler–Shi dataset. As additional comparison, we evaluate the uncertainty information that can be extracted from existing illuminant estimation methods by exploiting the natively-available confidence maps and weight maps (e.g., [10,27]). The experiment (reported in Appendix A.5) shows a very weak correlation, leading to the conclusion that such methods do not estimate uncertainty, even implicitly.

For each illuminant estimation method reported in Table 2 we select the uncertainty estimator having the highest correlation with the angular error in terms of PCC, and show in Fig. 2 the five images having the highest uncertainty. We observe that the majority of the reported images contain large objects with uniform color, which creates ambiguity in distinguishing between the contributions of the illuminant and surface reflectance. Additionally, some of these images belong to the "hard images" category identified by [26], and others contain multiple illuminants. In the top right corner of each image it is also reported the worst quantile to which the angular error belongs to (e.g., W-1% means that the image belongs to the first percentile of highest angular errors for a given method). We can observe how generally the images belong to the percentiles with largest illuminant estimation errors.

Since the numbers reported in Table 2 do not capture the complete behavior of the predicted uncertainties, we also report the following visualization in Fig. 3: given an illuminant estimation method and an uncertainty estimation method, we sort in increasing order the estimated uncertainties and group them in ten equal frequency bins. The illuminant estimation errors associated to the uncertainties in each bin are then represented with a box plot, showing the 5th and 95th percentiles, the 1st and 3rd quartiles, and the median value. Visualizations for a sample of the methods are reported in Fig. 3, while the others are provided in Appendix A.2. From the plot we can observe how, on average, the uncertainty associated to images with low illuminant estimation error is also low, while it increases as the illuminant estimation error increases. In general, the variance of illuminant estimation errors of low-uncertainty images is low, while it increases as the uncertainty increases. From the combination of the previous observations we can conclude that if the uncertainty is low, we can be confident that on that image the illuminant estimation error is also low. On the other hand, if the uncertainty is high then we cannot be sure if the illuminant estimation error is high or low, which is the definition of uncertainty itself.

In addition to the numerical results and the box plots reported in this section, in Appendix A.4 we also objectively evaluate the practical usability of the different uncertainty estimations.

**Table 2**
Results on Gehler–Shi dataset [33] with REC groundtruth [34]. For each column the best result is reported in bold. For each correlation type the best result per method is underlined.

| Method | Angular error [degrees] | | | | | Correlation (Pearson PCC) | | | | | Correlation (Spearman SRCC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | MC | Aleat. | Epist. | Intrinsic | | MC | Aleat. | Epist. | Intrinsic | |
| | Mean | Med. | Trim. | Q.-95 | Q.-99 | | | | Sing.S | Dual.S | | | | Sing.S. | Dual.S. |
| Cheng et al. [7] | 2.49 | 1.53 | 1.75 | 8.34 | 13.19 | – | .2986 | .4067 | – | .2330 | – | .3478 | .4945 | – | .3482 |
| FC$^4$ [10] | **2.14** | **1.44** | **1.57** | 6.50 | 12.75 | .4599 | .2969 | .3672 | – | .1932 | .4568 | .3356 | .4580 | – | .2487 |
| FC$^4$ [10] +SS | 2.30 | 1.62 | 1.80 | **6.48** | **10.87** | **.5331** | .3513 | .4678 | .4980 | – | **.5771** | **.4572** | .5629 | .5464 | – |
| Conv. Mean [36] | 2.50 | 1.73 | 1.90 | 7.93 | 12.42 | – | .3166 | .3630 | – | .2803 | – | .3814 | .4323 | – | .3686 |
| Conv. Mean [36] +SS | 2.95 | 1.90 | 2.21 | 8.78 | 14.17 | – | .3266 | .6170 | **.6293** | – | – | .3707 | **.6273** | **.5825** | – |
| Corr. Mom. [37] | 2.84 | 2.00 | 2.23 | 7.56 | 12.38 | – | .1612 | .4391 | – | .3153 | – | .2547 | .4922 | – | .3281 |
| Corr. Mom. [37] +SS | 3.33 | 2.14 | 2.33 | 8.23 | 16.56 | – | **.3885** | **.6969** | .5337 | – | – | .4208 | .5836 | .3514 | – |
| LMS Comm. [38] | 3.13 | 2.43 | 2.57 | 7.72 | 12.18 | – | .2254 | .4146 | – | **.4080** | – | .3512 | .5354 | – | **.4170** |

**Table 3**
Results on Simple Cube++ dataset [35]. For each column the best result is reported in bold. For each correlation type the best result per method is underlined.

| Method | Angular error [degrees] | | | | | Correlation (Pearson PCC) | | | | | Correlation (Spearman SRCC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | MC | Aleat. | Epist. | Intrinsic | | MC | Aleat. | Epist. | Intrinsic | |
| | Mean | Med. | Trim. | Q.-95 | Q.-99 | | | | Sing.S. | Dual.S. | | | | Sing.S. | Dual.S. |
| Cheng et al. [7] | 1.39 | 0.63 | 0.72 | 5.44 | 12.61 | – | .2782 | .4731 | – | .3850 | – | .3109 | .3527 | – | .4139 |
| FC$^4$ [10] | **1.11** | **0.58** | **0.66** | **3.82** | 9.59 | .4725 | .3333 | .4503 | – | .4424 | .4887 | .3252 | .4606 | – | .4184 |
| FC$^4$ [10] + SS | 1.38 | 0.70 | 0.85 | 4.40 | 9.80 | **.6452** | .4001 | .6272 | .6316 | – | **.6324** | .4049 | .6152 | **.6430** | – |
| Conv. Mean [36] | 1.53 | 0.76 | 0.91 | 5.67 | 11.47 | – | .3503 | .5072 | – | **.4941** | – | .4228 | .5286 | – | **.4568** |
| Conv. Mean [36] + SS | 1.90 | 1.03 | 1.31 | 6.03 | 9.59 | – | .5704 | .6803 | .6226 | – | – | **.5833** | .7056 | .5606 | – |
| Corr. Mom. [37] | 2.00 | 1.32 | 1.47 | 6.39 | **9.20** | – | .4024 | .5833 | – | .3338 | – | .4608 | .6899 | – | .3870 |
| Corr. Mom. [37] +SS | 2.23 | 1.35 | 1.55 | 6.70 | 14.29 | – | **.6306** | **.6928** | **.7366** | – | – | .5768 | **.7315** | .5773 | – |
| LMS Comm. [38] | 2.38 | 1.77 | 1.83 | 6.58 | 12.94 | – | .3021 | .6776 | – | .2889 | – | .3329 | .6923 | – | .2571 |


**Fig. 2.** Top five images on which each method has the highest uncertainty (in decreasing order). In the top right corner of each image it is reported the worst percentile to which the angular error of the illuminant estimation algorithm applied to the specific image belongs to.

### 3.1. Intrinsic uncertainty hyperparameters optimization

The weights $(\lambda_{IE}, \lambda_{UE\text{-}L1}, \lambda_{UE\text{-}C})$ used in all experiments regarding intrinsic uncertainty have been selected by hyperparameter optimization using FC$^4$ as reference model (see Section 2.1). Here we plot the Pareto front in terms of median angular error vs. PCC that can be achieved by sweeping over the weight parameters. The considered methods are FC$^4$, Convolutional Mean, and Corrected Moments. Due to the different training times, the Pareto front for Corrected Moments is the most dense. The plots are depicted in Fig. 4, where we can observe a similar behavior of the three intrinsic SS methods: at low angular errors the plots have a very high slope, meaning that by allowing a slight increase in the angular error we can have a large improvement in correlation; at high angular errors we have low slopes, meaning that to obtain an improvement at high correlation values we have to allow a large increase in the angular error. Overall we can observe how FC$^4$ can reach the lowest angular errors, while Convolutional Mean reaches

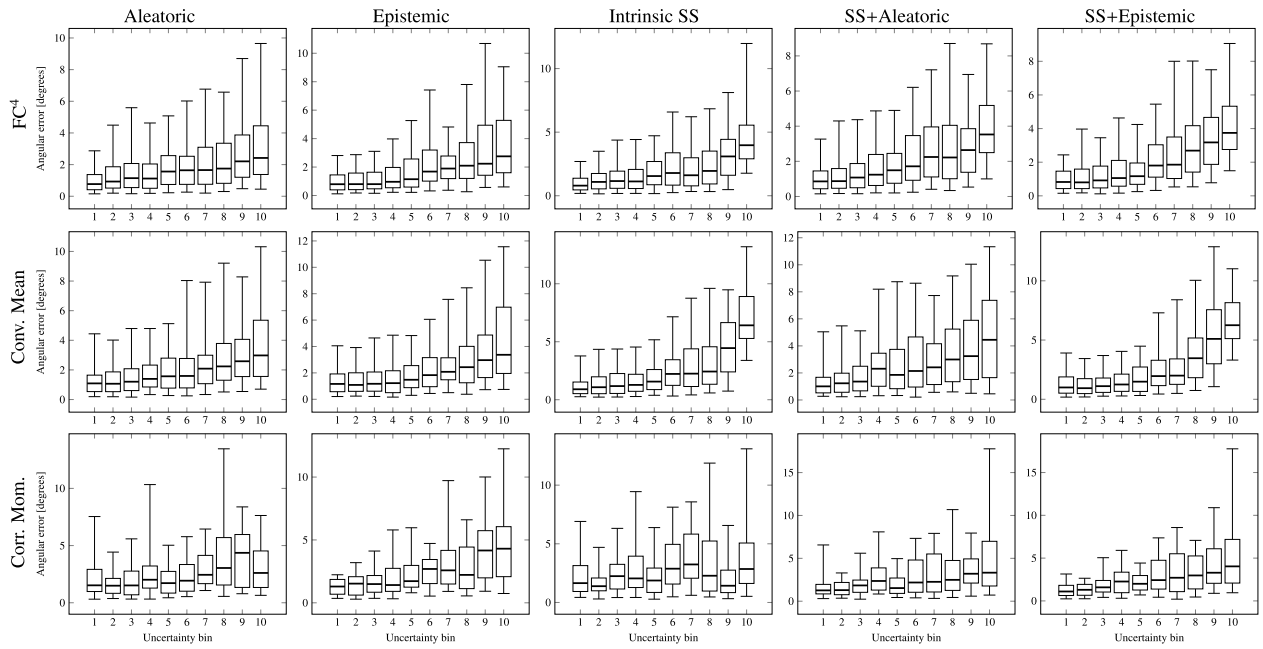


**Fig. 3.** Uncertainty visualization. Uncertainty is computed on the Gehler–Shi dataset, sorted in increasing order and grouped in ten equal frequency bins. The illuminant estimation errors on the images falling in each bin are represented with a boxplot. Three illuminant estimation algorithms are reported (on the rows: FC$^4$ [10], Convolutional Mean [36] and Corrected Moments [37]) and five different uncertainties (on the columns: aleatoric, epistemic, intrinsic SS, SS with aleatoric, SS with epistemic).

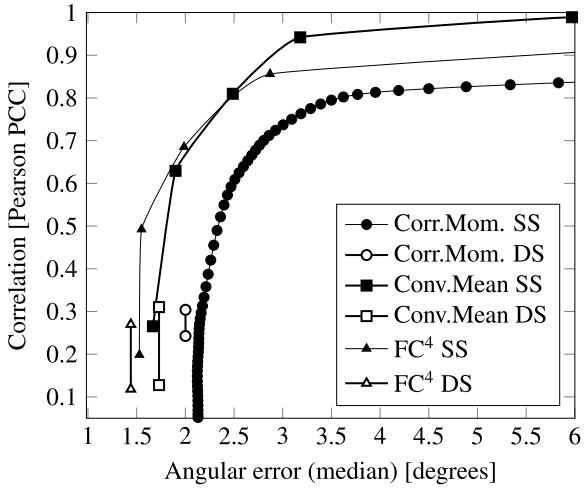


**Fig. 4.** Pareto front for Corrected [37], Convolutional Mean [36] and FC$^4$ [10] generated by sweeping the weights in Eq. (16).

the highest correlations. Remarkably, Convolutional Mean is the only method for which SS is able to reach lower angular errors than DS.

### 3.2. Ablation study

In this section we run an ablation study on the total loss $\mathcal{L}_{TOT}$ reported in Eq. (16). In all the configurations tested the term $\mathcal{L}_{IE}$ is always present, since it is the loss commonly used for training illuminant estimation methods. The weights $\lambda_*$ of the different loss components are the same described in Section 2.1. The results of the ablation study reported in Table 4 show that using only the illuminant estimation $\mathcal{L}_{IE}$ term, the method reaches the lowest average, median, and trimean angular errors and a poor uncertainty estimation with an average PCC correlation of 0.3293 and an average SRCC correlation of 0.3748. Adding the $\mathcal{L}_{UE-L1}$ term in the total loss slightly increases the average, median, and trimean angular errors, while slightly reducing the 99th percentile of the angular error; also the average PCC and SRCC correlations are slightly reduced to 0.3081 and 0.3708 respectively. Replacing the $\mathcal{L}_{UE-L1}$ term with the $\mathcal{L}_{UE-C}$ term instead, increases the average, median, and trimean angular errors, and reduces the 99th percentile of the angular error; the use of this terms shows a great boost in correlation, with an average PCC correlation of 0.4501 and an average SRCC correlation of 0.5224. Adding the $\mathcal{L}_{UE-L1}$ in the total loss, i.e. considering all the proposed terms, slightly reduces the average, median, and trimean angular errors with respect to the previous configuration, and obtains the lowest 95th and 99th percentiles of the angular error; also the quality of the estimated uncertainty increases, with an average PCC correlation of 0.4626 and an average SRCC correlation of 0.5359.

### 3.3. Uncertainty-based cascaded color constancy

In this section we conduct a preliminary experiment providing two use cases to show how image processing pipelines can leverage the estimated uncertainty. In the first use case we assume to have two different illuminant estimation algorithms: the first one is less accurate but it has a lower count of floating point operations (FLOPs); the second one is more accurate but also has a higher FLOPs count. Suppose that the considered image processing pipeline has a maximum limit on the FLOPs per image, which prevents using the most expensive algorithms on all images/frames. In this configuration we can use the first algorithm, and only if its uncertainty on the estimate is higher than a given threshold $t_u$ we run the second algorithm and use its estimate. The result of such experiment is reported in Fig. 5, where we vary $t_u$ and report the angular error (mean, median, and Q.-95) and the average number of FLOPs per image on the Gehler–Shi dataset. In this experiment Conv.Mean is used as the first algorithm (approx. 0.05 G-FLOPs), and FC$^4$ as the second one (approx. 1.65 G-FLOPs). For both algorithms the uncertainty used is the SS intrinsic uncertainty. From the plot we can observe how the three error statistics decrease as the number of G-FLOPs increases. Furthermore we can observe how this error reduction is highly nonlinear for the 95th-percentile, with a reduction of about one degree at the cost of 0.26 more G-FLOPs, while for a reduction of a further one degree, 0.89 additional G-FLOPs are needed.

**Table 4**
Ablation study on the total loss $\mathcal{L}_{TOT}$ for the FC$^4$ [10] illuminant estimation method on Gehler–Shi dataset [33] with REC groundtruth [34]. For each column the best result is reported in bold. For each correlation type the best result per method is underlined.

| Method | Loss terms | | | Angular error [degrees] | | | | | Correlation (Pearson PCC) | | | | | Correlation (Spearman SRCC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | MC | Aleat. | Epist. | Intrinsic | | MC | Aleat. | Epist. | Intrinsic | |
| | $\mathcal{L}_{IE}$ | $\mathcal{L}_{UE-L1}$ | $\mathcal{L}_{UE-C}$ | Mean | Med. | Trim. | Q.-95 | Q.-99 | | | | Sing.S | Dual.S | | | | Sing.S. | Dual.S. |
| FC$^4$ [10] | ✓ | | | **2.14** | **1.44** | **1.57** | 6.50 | 12.75 | .4599 | .2969 | .3672 | – | .1932 | .4568 | .3356 | .4580 | – | .2487 |
| FC$^4$ [10] +SS | ✓ | ✓ | | 2.21 | 1.47 | 1.65 | 6.77 | 12.10 | .4408 | .2571 | .3649 | .1694 | – | .4628 | .3132 | .4692 | .2379 | – |
| FC$^4$ [10] +SS | ✓ | | ✓ | 2.32 | 1.65 | 1.80 | 6.75 | 11.38 | .5205 | .2952 | **.4711** | **.5134** | – | **.5819** | .3800 | .5624 | **.5652** | – |
| FC$^4$ [10] +SS | ✓ | ✓ | ✓ | 2.30 | 1.62 | 1.80 | **6.48** | **10.87** | **.5331** | **.3513** | .4678 | .4980 | – | .5771 | **.4572** | **.5629** | .5464 | – |

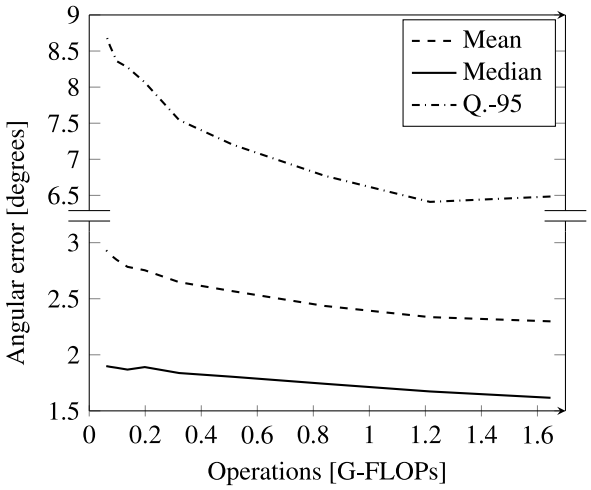


**Fig. 5.** Performance of uncertainty-based cascaded illuminant estimation with a limit on the G-FLOPs per image/frame.

In the second use case we are only interested in performing the most accurate illuminant estimate. We start from FC$^4$ with SS intrinsic uncertainty and we investigate if another algorithm, i.e. Conv.Mean with SS intrinsic uncertainty, can help improving the estimation when the first uncertainty is high. Let $\hat{y}_1$ be the illuminant estimated by the first algorithm and $u_1$ its estimated uncertainty; let $\hat{y}_2$ and $u_2$ be the corresponding estimates of the second algorithm. If $u_1 \leq t_1$, we use the estimate $\hat{y}_1$ given by the first algorithm. If $u_1 > t_1$, we also run the second algorithm: if $u_2 < u_1$, i.e. the second algorithm is less uncertain than the first one on the given image, we use the estimate $\hat{y}_2$ given by the second algorithm, otherwise we use $\hat{y}_1$. The performance of this cascaded illuminant estimation is reported in terms of angular error in Table 5, where for sake of comparison we also report the individual performance of the two algorithms used in the cascade. From the values reported in Table 5 it is possible to notice how the uncertainty-based cascade algorithm is able to improve on all the error statistics with respect to the best method used in the cascade, with the highest improvement on the higher percentiles. The second algorithm is run on about 61% of the images, and its estimate is actually used in 25% of the cases, resulting in an average of 1.68 G-FLOPs per image.

Another way in which the estimated uncertainty could be used is for enlarging color constancy datasets by providing pseudo-labels for unlabeled images by considering only the predictions with low uncertainty [43].

## 4. Conclusion

In this paper we presented a formalization of uncertainty estimation in color constancy, and we defined three forms of uncertainty that require at most one inference run to be estimated, opposed to existing methods where several runs are needed (e.g. more than 10). We applied the defined uncertainty estimators to five different categories of color constancy algorithms: tree-based, deep CNN-based, shallow CNN-based, statistics-based, and combination-based. Experimental results on two standard color constancy datasets showed a strong correlation between the estimated uncertainty and the illuminant estimation error; in contrast, the uncertainty information that can be extracted from illuminant estimation methods natively exploiting confidence maps and weight maps showed at most just a weak correlation with the illuminant estimation error. Furthermore, we showed two possible uses cases of how color constancy algorithms can be cascaded leveraging the estimated uncertainty to provide more accurate illuminant estimates.

For the purpose of this paper, we focused on uncertainty in single-frame single-illuminant RGB illuminant estimation. Additional challenges might be found in a variety of domain extensions, namely: temporal, spectral, and spatial extensions. Nonetheless, we consider it important to lay the groundwork for uncertainty estimation in color constancy at its most widespread basic interpretation. This, in our opinion, will bootstrap the research on the aforementioned extensions, which we in fact intend to consider for future developments.

As future work we also plan to apply the defined uncertainty estimators to a larger set of color constancy algorithms. Finally, as a further research direction we plan to exploit the uncertainty information to create labels for unlabeled images, thus enlarging the size of color constancy datasets and also permitting to design new color constancy algorithms exploiting different learning paradigms, as for example semi-supervised learning.

**Table 5**
Angular error statistics on Gehler–Shi dataset [33] with REC groundtruth [34] for the uncertainty-based cascaded illuminant estimation compared with the two cascaded algorithms.

| Method | Angular error [degrees] | | | | |
|---|---|---|---|---|---|
| | Mean | Med. | Trim. | Q.-95 | Q.-99 |
| FC$^4$ [10] + SS | 2.30 | 1.62 | 1.80 | 6.48 | 10.87 |
| Conv. Mean [36] + SS | 2.95 | 1.90 | 2.21 | 8.78 | 14.17 |
| Unc.-based Cascaded IE | **2.22** | **1.55** | **1.73** | **6.15** | **9.86** |

## CRediT authorship contribution statement

**Marco Buzzelli:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Simone Bianco:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

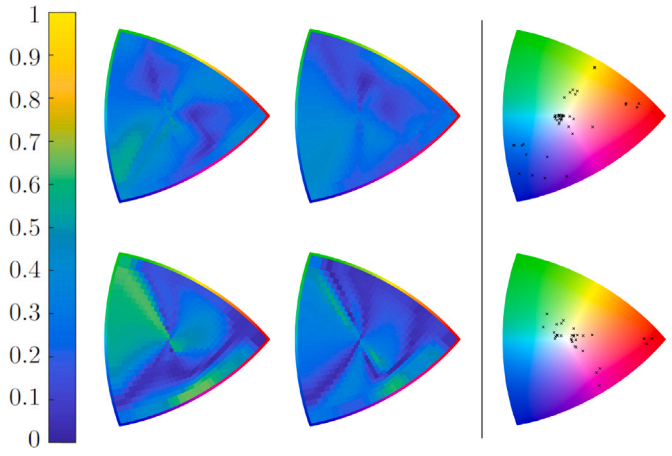


**Fig. A.6.** Aleatoric uncertainty. Left and center: correlation (absolute Pearson and Spearman) obtained with all 629 illuminant perturbations for the Corrected Moments + SS illuminant estimation method. Right: best uncertainty predictors across different methods. Top: Gehler–Shi dataset [33]. Bottom: SimpleCube++ dataset [35]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

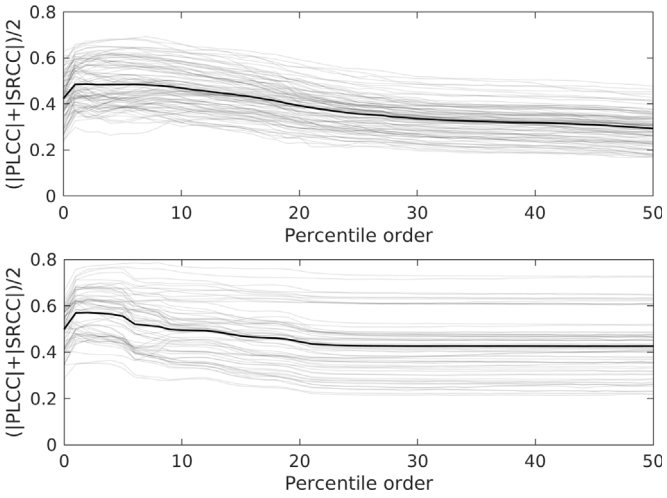


**Fig. A.7.** Epistemic uncertainty correlation (average of absolute Pearson and Spearman) obtained with different percentile orders across different methods. Top: Gehler–Shi dataset [33]. Bottom: SimpleCube++ dataset [35].

## Appendix

### A.1. Best aleatoric and epistemic predictors

Fig. A.6 presents a visualization of the data underlying aleatoric uncertainty estimation. On the left and center we focus on the method that produces the best uncertainty estimation: Corrected Moments [37] adapted for single shot intrinsic uncertainty. Here, a heatmap is used to visualize the correlation (absolute Pearson and Spearman respectively) obtained using all 629 considered illuminant perturbations, for the two dataset (Gehler–Shi in the top, SimpleCube++ in the bottom). This visualization highlights the elevated variability of responses within the chromaticity diagram, showing that a small change might lead to a drastic reduction in uncertainty estimation performance. On the right, we document the best predictive perturbations for the two datasets, selected across all analyzed illuminant estimation methods for different folds and runs. Both datasets present a higher density around less saturated illuminants (closer to the center), with Gehler–Shi predominantly in the cyan area, and SimpleCube++ split between cyan and red.

In Fig. A.7 we report the data underlying epistemic uncertainty, by plotting the different correlation values (average of absolute Pearson and Spearman) obtained with percentile orders between 0 and 50. Different illuminant estimation methods, folds and runs are overimposed and averaged. The monotonically-decreasing trend shows that selecting one of the closest ground truth illuminants is a good predictor for uncertainty. We can also observe how a non-zero-order percentile is a better candidate than the minimum distance, providing a more robust selection.

### A.2. Uncertainty visualizations

In this section we report the box plot visualizations for the remaining combinations of illuminant estimation algorithm and uncertainty type that are not reported in the main document: MC dropout on $FC^4$ and $FC^4$+SS (Fig. A.8); Aleatoric uncertainty on [7] and LMS Committee [38] (Fig. A.9); Epistemic uncertainty on [7] and LMS Committee [38] (Fig. A.10); Dual Shot Intrinsic uncertainty (Fig. A.11). For each combination, we sort in increasing order the estimated uncertainties and group them in ten equal frequency bins.

### A.3. Performance analysis of MC dropout

In Fig. A.12 we analyze the performance of MC dropout [22] by varying the number of stochastic forward passes $T$ through the trained model in the range $2 \leq T \leq 100$, $T \in \mathbb{N}$. Recall that in the comparisons in the main document we use $T = 100$.

### A.4. Uncertainty usability

In addition to the numerical results and the box plots reported in Section 3 of the paper, in this section we aim to objectively evaluate the practical usability of the different uncertainty estimations. We define an uncertainty estimate to be usable if the following three conditions hold: (i) highest uncertainty values are assigned to images on which the illuminant estimation method produces the highest errors. (ii) the standard deviation of the illuminant estimation errors is small, i.e. images are assigned uncertainty values that correlate well with the illuminant estimation errors. A sub-optimal configuration takes place if the standard deviation increases for increasing uncertainty values, i.e. when a low uncertainty value is assigned we are sure that the illuminant estimation error is low, but when a high uncertainty value is assigned we cannot know if the illuminant estimation error is high or low. (iii) the uncertainty value is close to the illuminant estimation error.

The first condition is quantified by computing the SRCC between the median values of the box plots similar to those reported in Figs. A.8, A.9, A.10, A.11 and a monotonically increasing sequence. The second condition is quantified by computing the SRCC between the difference of the 95th and 5th quantiles (i.e. the whiskers) of the previous box plots and a monotonically increasing sequence. Finally, the third condition is quantified by computing the Mean Absolute Error (MAE) between the uncertainty value and the illuminant estimation error. The computed values are reported in Table A.6. From the reported numbers we can observe how the MC dropout, epistemic and intrinsic SS uncertainty reach the highest usability, confirming the qualitative analysis from Section 3 of the paper. Furthermore, we can observe once again the improvement that is gained when epistemic uncertainty is applied to an illuminant estimation method trained with intrinsic SS uncertainty.

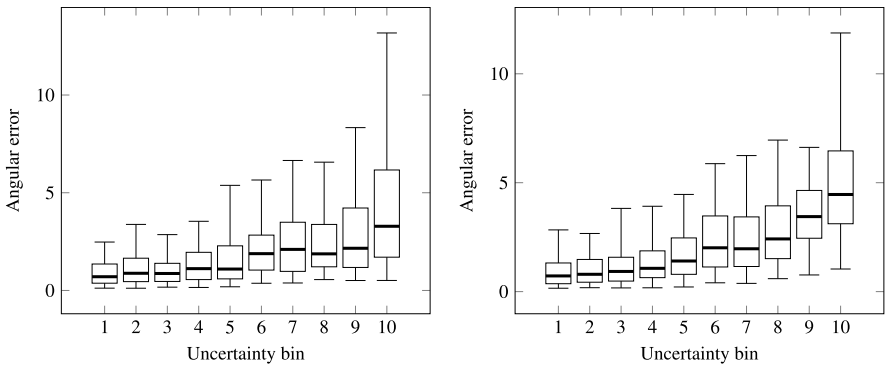


**Fig. A.8.** MC dropout uncertainty [22] visualization: FC⁴ [10] (left), and FC⁴+SS (right).

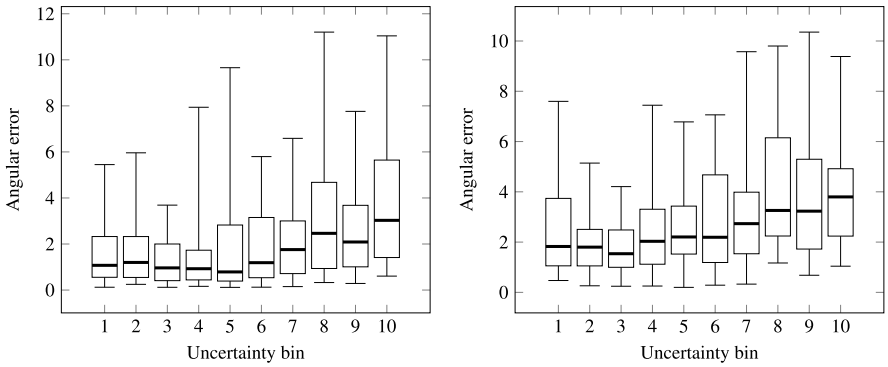


**Fig. A.9.** Aleatoric uncertainty visualization: [7] (left), and LMS Committee [38] (right).

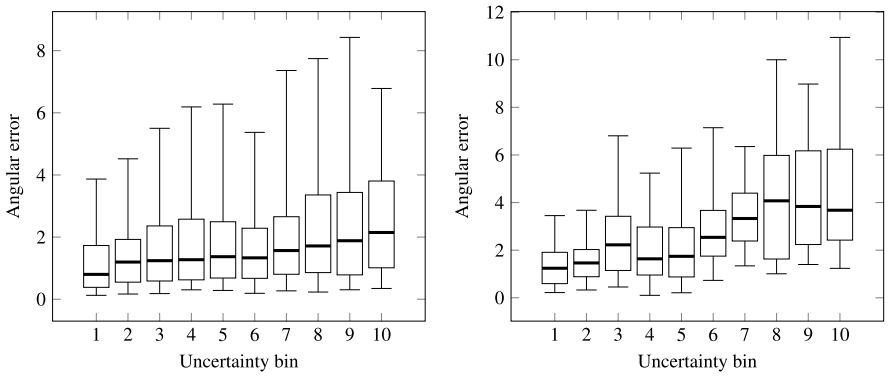


**Fig. A.10.** Epistemic uncertainty visualization: [7] (left), and LMS Committee [38] (right).

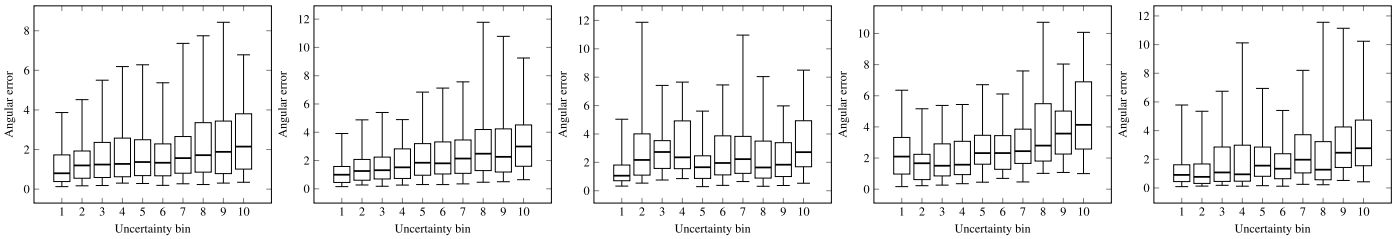


**Fig. A.11.** Double Shot Intrinsic uncertainty visualization. Left to right: FC⁴ [10], Convolutional Mean [36], Corrected Moments [37], LMS Committee [38], and [7].

### *A.5. Do existing color constancy methods already implicitly estimate uncertainty?*

In this section we aim to understand if existing illuminant estimation algorithms that employ the concepts of confidence-weighted pooling [10] or weight maps [27] already implicitly quantify the uncertainty associated to each estimate. To this end, we compute several punctual statistics (i.e. average, median, standard deviation, 95th quantile, and maximum) from the confidence maps and weight maps, and we compute their Pearson and Spearman correlation with the angular error. The results are reported in Table A.7 on the Gehler–Shi dataset [33]. In the same table, several statistics on the angular

**Table A.6**
Quantitative estimation of uncertainty usability computed on the Gehler–Shi dataset [33] with REC groundtruth [34]: SRCC between the median values of the boxplots and a monotonically increasing sequence (M); SRCC between the standard deviations of the bins in the boxplots and a monotonically increasing sequence (W); Mean Absolute Error (MAE) between the uncertainty values and the illuminant estimation error. For each column the best value is reported in bold. For each row the best value for each measure is underlined.

| Method | MC dropout | | | Aleatoric | | | Epistemic | | | Intrinsic Single Shot | | | Intrinsic Dual Shot | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | W | MAE | M | W | MAE | M | W | MAE | M | W | MAE | M | W | MAE |
| [7] | – | – | – | 0.709 | 0.673 | 2.08 | 0.891 | 0.939 | 1.95 | – | – | – | 0.891 | 0.685 | 1.84 |
| FC$^4$ [10] | 0.939 | **0.976** | 1.58 | **0.988** | **0.952** | 1.43 | **1.000** | 0.952 | 1.40 | – | – | – | **0.988** | 0.855 | **1.48** |
| FC$^4$ [10] +SS | **0.988** | 0.952 | **1.27** | **0.988** | 0.939 | **1.42** | 0.988 | 0.939 | **1.34** | 0.976 | **0.988** | **1.45** | – | – | – |
| Conv. Mean [36] | – | – | – | **0.988** | 0.915 | 1.59 | 0.988 | **0.976** | 1.57 | – | – | – | 0.976 | **0.939** | 1.62 |
| Conv. Mean [36] +SS | – | – | – | 0.964 | 0.915 | 1.95 | 0.988 | 0.891 | 1.54 | **1.000** | **0.988** | 1.55 | – | – | – |
| Corr. Mom. [37] | – | – | – | 0.891 | 0.358 | 2.16 | 0.903 | 0.939 | 1.82 | – | – | – | 0.103 | 0.273 | 2.40 |
| Corr. Mom. [37] +SS | – | – | – | 0.903 | 0.818 | 2.65 | 0.988 | 0.915 | 2.47 | 0.467 | 0.612 | 3.03 | – | – | – |
| LMS Comm. [38] | – | – | – | 0.927 | 0.721 | 1.79 | 0.915 | 0.830 | 1.66 | – | – | – | 0.891 | 0.782 | 1.70 |

**Table A.7**
Results on Gehler–Shi dataset [33] with REC groundtruth [34]. For each column the best result is reported in bold. For each correlation type the best result per method is underlined. Correlation is computed between the illuminant estimation error and several statistics (Mean, Median, Standard deviation, 95th quantile and Maximum) extracted from the confidence map [10] and the weight map [27].

| Method | Angular error [degrees] | | | | | Correlation (Pearson PCC) | | | | | Correlation (Spearman SRCC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Med. | Trim. | Q.-95 | Q.-99 | Mean | Med. | Std | Q.-95 | Max | Mean | Med. | Std. | Q.-95 | Max |
| FC$^4$ [10] | **2.14** | **1.44** | **1.57** | **6.50** | **12.75** | **.1101** | **.1205** | .0933 | .1019 | .0992 | .1616 | **.1646** | .1390 | .1456 | **.1459** |
| QU [27] | 3.19 | 1.99 | 2.32 | 10.52 | 13.59 | .0922 | .0585 | **.1746** | **.1893** | **.2201** | **.2113** | .0776 | **.2340** | **.2273** | .1452 |

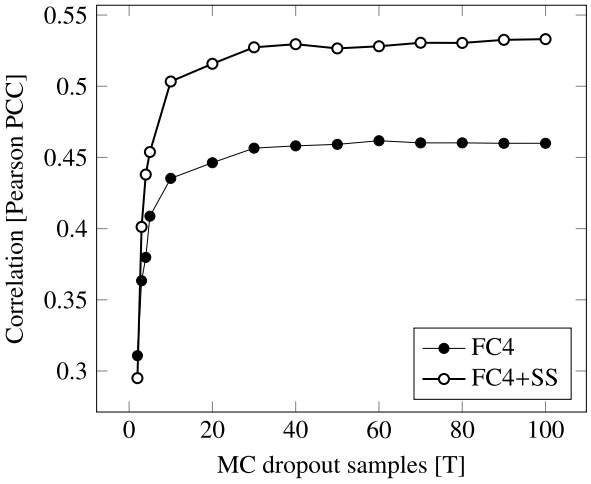


**Fig. A.12.** Performance of MC dropout [22] by varying the number of stochastic forward passes $T$ through the trained model.

errors are also reported. The values reported in Table A.7 show that for both methods the correlation is just weak or even very weak. We can therefore conclude that these methods do not natively associate an uncertainty value to their illuminant estimates.

## Data availability

No data was used for the research described in the article.

## References


[1] B. Funt, K. Barnard, L. Martin, Is machine colour constancy good enough? in: Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June, 2–6, 1998 Proceedings, Volume I 5, Springer, 1998, pp. 445–459.

[2] C. Witzel, C. van Alphen, C. Godau, J.K. O'Regan, Uncertainty of sensory signal explains variation of color constancy, J. Vis. 16 (15) (2016) 8.

[3] D.H. Foster, A. Reeves, Colour constancy failures expected in colourful environments, Proc. R. Soc. Lond. [Biol.] 289 (1967) (2022) 20212483.

[4] J. Van De Weijer, T. Gevers, A. Gijsenij, Edge-based color constancy, IEEE Trans. Image Process. 16 (9) (2007) 2207–2214.

[5] V.C. Cardei, B. Funt, K. Barnard, Estimating the scene illumination chromaticity by using a neural network, J. Opt. Soc. Amer. A 19 (12) (2002) 2374–2386.

[6] W. Xiong, B. Funt, Estimating illumination chromaticity via support vector regression, J. Imaging Sci. Technol. 50 (4) (2006) 341–348.

[7] D. Cheng, B. Price, S. Cohen, M.S. Brown, Effective learning-based illuminant estimation using simple features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1000–1008.

[8] S.W. Oh, S.J. Kim, Approaching the computational color constancy as a classification problem through deep learning, Pattern Recognit. 61 (2017) 405–416.

[9] W. Shi, C.C. Loy, X. Tang, Deep specialized network for illuminant estimation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 371–387.

[10] Y. Hu, B. Wang, S. Lin, Fc4: Fully convolutional color constancy with confidence-weighted pooling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4085–4094.

[11] S. Bianco, C. Cusano, R. Schettini, Single and multiple illuminant estimation using convolutional neural networks, IEEE Trans. Image Process. 26 (9) (2017) 4347–4362.

[12] T. Tommasi, N. Patricia, B. Caputo, T. Tuytelaars, A deeper look at dataset bias, Domain Adapt. Comput. Vis. Appl. (2017) 37–55.

[13] M. Buzzelli, S. Zini, S. Bianco, G. Ciocca, R. Schettini, M.K. Tchobanou, Analysis of biases in automatic white balance datasets and methods, Color Res. Appl. 48 (1) (2023) 40–62.

[14] P. Morovič, J. Morovič, Atomic color: From points to probability distributions, in: Computational Color Imaging: 8th International Workshop, CCIW 2024, Milan, Italy, September 25-27, 2024, Proceedings 8, Springer, 2024.

[15] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[16] S. Zini, A. Gomez-Villa, M. Buzzelli, B. Twardowski, A.D. Bagdanov, J. Van De Weijer, Planckian jitter: Countering the color-crippling effects of color jitter on self-supervised training, in: The Eleventh International Conference on Learning Representations, 2023, URL https://openreview.net/forum?id=Pia70sP2Oi1.

[17] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2) (2009) 105–112.

[18] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30 (2017).

[19] L. Nanni, S. Ghidoni, S. Brahnam, Handcrafted vs. non-handcrafted features for computer vision classification, Pattern Recognit. 71 (2017) 158–172.

[20] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, et al., Deep learning for computer vision: A brief review, Comput. Intell. Neurosci. 2018 (2018).

[21] J.M. Hernández-Lobato, R. Adams, Probabilistic backpropagation for scalable learning of bayesian neural networks, in: International Conference on Machine Learning, PMLR, 2015, pp. 1861–1869.

[22] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[23] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Adv. Neural Inf. Process. Syst. 30 (2017).
[24] A. Loquercio, M. Segu, D. Scaramuzza, A general framework for uncertainty estimation in deep learning, IEEE Robot. Autom. Lett. 5 (2) (2020) 3153–3160.
[25] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, Adv. Neural Inf. Process. Syst. 32 (2019).
[26] R. Zakizadeh, M.S. Brown, G.D. Finlayson, A hybrid strategy for illuminant estimation targeting hard images, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 16–23.
[27] S. Bianco, C. Cusano, Quasi-unsupervised color constancy, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12212–12221.
[28] J.T. Barron, Convolutional color constancy, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 379–387.
[29] S.D. Hordley, G.D. Finlayson, Reevaluation of color constancy algorithm performance, J. Opt. Soc. Amer. A 23 (5) (2006) 1008–1020.
[30] D.L. MacAdam, Sources of Color Science, MIT Press, 1970.
[31] Y. Dodge, The Concise Encyclopedia of Statistics, Springer New York, 2008.
[32] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. (2021).
[33] P.V. Gehler, C. Rother, A. Blake, T. Minka, T. Sharp, Bayesian color constancy revisited, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[34] G. Hemrit, G.D. Finlayson, A. Gijsenij, P. Gehler, S. Bianco, M.S. Drew, B. Funt, L. Shi, Providing a single ground-truth for illuminant estimation for the ColorChecker dataset, IEEE Trans. Pattern Anal. Mach. Intell. 42 (5) (2019) 1286–1287.
[35] E. Ershov, A. Savchik, I. Semenkov, N. Banić, A. Belokopytov, D. Senshina, K. Koščević, M. Subašić, S. Lončarić, The cube++ illumination estimation dataset, IEEE Access 8 (2020) 227511–227527.
[36] H. Gong, Convolutional mean: A simple convolutional neural network for illuminant estimation, in: 30th British Machine Vision Conference 2019, BMVC 2019, 2020.
[37] G.D. Finlayson, Corrected-moment illuminant estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1904–1911.
[38] V.C. Cardei, B. Funt, Committee-based color constancy, in: Color and Imaging Conference, Society for Imaging Science and Technology, 1999, pp. 311–313.
[39] S. Zini, M. Buzzelli, S. Bianco, R. Schettini, COCOA: Combining color constancy algorithms for images and videos, IEEE Trans. Comput. Imaging 8 (2022) 795–807.
[40] M. Buzzelli, S. Bianco, R. Schettini, ARC: Angle-retaining chromaticity diagram for color constancy error analysis, J. Opt. Soc. Amer. A 37 (11) (2020) 1721–1730.
[41] S.H. Yelleni, D. Kumari, S.P. K., K.M. C., Monte Carlo DropBlock for modeling uncertainty in object detection, Pattern Recognit. 146 (2024) 110003.
[42] H. Weisberg, Central Tendency and Variability, (83) Sage, 1992.
[43] S. Kim, P. Chikontwe, S. An, S.H. Park, Uncertainty-aware semi-supervised few shot segmentation, Pattern Recognit. 137 (2023) 109292.

**Marco Buzzelli** obtained his Ph.D. in Computer Science in 2019 at the University of Milano–Bicocca (Italy), where he currently serves as assistant professor at the Department of Informatics, Systems and Communication. He is actively engaged in conducting cutting-edge research in the field of signal/image/video processing and understanding, using machine learning techniques. He is particularly passionate about color imaging. Marco Buzzelli has actively collaborated with European institutions, including but not limited to Universitat Autoònoma de Barcelona, Universidade Nova de Lisboa, Université Jean Monnet, Universidad de Granada. These collaborations have allowed him to contribute to the European AI landscape as an active ELLIS member, while also gaining valuable insights and exposure to diverse perspectives.

**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).