Pattern Recognition Letters
journal homepage: www.elsevier.com

Camera-Independent Color Constancy by Scene Semantics

Mengda Xie[a, b], Peng Sun[a] *, Yubo Lang[a] and Meie Fang[b]

[a]*Department of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang 110035, China*
[b]*Department of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 511400, China*

ABSTRACT

Current learning-based color constancy methods are typically employed to find camera-specific illuminant mappings. Consequently, these methods exhibit poor generalization to images captured by varying cameras. In this paper, we present a Camera-Independent learning method based on Scene Semantics, and we call it CISS. Inspired by the camera-independent property of gray-based methods, CISS does not directly estimate camera-specific illuminant by training model as most learning methods do. Instead, the model's output is transformed into camera-independent scene statistics related to gray-based assumptions to avoid being affected by camera variations. Based on these estimated scene statistics, illuminant can be calculated indirectly. To estimate scene statistics accurately, CISS designs illuminant-invariance scene semantics features as input to the model. Then, the model estimates scene statistics for each input image in terms of scene semantics with exemplar-based learning. Experiments show that, on several public datasets, CISS is able to outperform present methods for multi-cameras color constancy, and is flexible enough to be well generalized to the unseen camera without fine-tuning by additional images.

Keywords: Color constancy, Illuminant estimation, Scene semantics, Gray-based assumptions

1. Introduction

The image value $\boldsymbol{I}=(I_R,I_G,I_B)^T$ depends on the camera sensor function $\boldsymbol{F}(\lambda)=(F_R(\lambda),F_G(\lambda),F_B(\lambda))^T$, the illuminant $L(\lambda)$ and the surface reflectance $S(x,y,\lambda)$ at pixel location (x,y) [1].

$$I_c(x,y)=\int_\omega F_c(\lambda)L(\lambda)S(x,y,\lambda)d\lambda \quad (1)$$

where $c=\{R,G,B\}$, ω is the visible spectrum. Assuming a unique illuminant in the scene, the observed illuminant $\boldsymbol{E}$ depends on the illuminant $L(\lambda)$ as well as the camera sensor function $\boldsymbol{F}(\lambda)$.

$$\boldsymbol{E}=\int_\omega L(\lambda)\boldsymbol{F}(\lambda)d\lambda \quad (2)$$

Given the image values of $\boldsymbol{I}$, color constancy is targeted for estimating the color of the illuminant $\boldsymbol{E}=(E_R,E_G,E_B)^T$, followed by a transformation of the color-biased image using this illuminant estimate. According to whether a training process is necessary, the most existing color constancy method can be roughly classified as learning-free methods and learning-based methods [2]. The learning-based methods [3-6], including deep neural networks (DNN) [7, 8], have exhibited significant outperformance on specific datasets. However, their training phase is clearly relied on the camera-specific illuminant (see Eq. 2) supplied by the dataset. Fig. 1 (a) visualizes the illuminant distributions from two datasets captured by different cameras. It is clear that these illuminant distributions change significantly across the two cameras because of their different spectral sensitivities. As a consequence, once these learning-based methods are well trained on a dataset taken by specific camera, they are hardly generalized to another dataset captured by different cameras. On the other hand, Learning-free methods, such as gray-based methods [9-12], generate fixed assumptions related to scene statistics for all images and performed illuminant estimation. Although these assumptions are less accurate [13, 14], they are generally invariant to much of the spectral sensitivity differences among camera sensors [2, 13]. As shown in Fig. 1 (b), under a neutral illuminant, the scene statistic distributions related to the gray world assumption are similar in images taken by different cameras. Due to this property, gray-based methods have the natural advantage of being camera-independent and, therefore, very well-suited to the multi-cameras task.

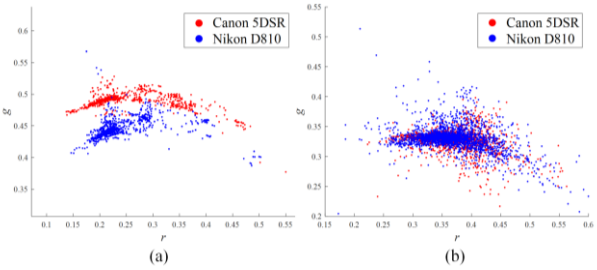


Fig. 1. (a) Under a neutral illuminant, the *rg*-chromaticities visualization distributions of the scene statistic related to the gray world assumption in images taken by different cameras. (b) The *rg*-chromaticities visualization distribution of illuminant in the same images.

* Corresponding author.
E-mail: 1112206011@e.gzhu.deu.cn (Mengda Xie), sunpeng_sx@cipuc.edu.cn (Peng Sun), langyubo@cipuc.edu.cn (Yubo Lang), fme@gzhu.edu.cn (Meie Fang)

In this paper, we present CISS, a camera-independent learning method based on scene semantics. Different from most learning-based methods that directly estimate the illuminant, we perform indirect illuminant estimation by calculating the scene statistics for each image under a neutral illuminant as new labels. Given that these scene statistics are camera-independent (see Fig. 1 (b)), the model will be less affected by camera changes during both the training and testing phases. Once the trained model gives the scene statistics of the input image, the illuminant of the input image can be calculated indirectly in a similar way to gray-based methods. In more detail, for accurately estimating the scene statistics of the color-biased input image when it is under a neutral illuminant, CISS designs illuminant-invariance cross-moments to extract robust local color features. Next, these local color features are used to further extract the scene semantics with spatial pyramid structure based on PLSA as well as spatial pyramid blocks. Subsequently, by means of exemplar-based learning, CISS selects nearest neighbor images with high scene semantics similarity for each input image, and calculates the scene statistics of these selected images when they are under a neutral illuminant. Finally, the median of these scene statistics is used as the scene statistic for the input image under neutral illuminant. Four contributions in this paper can be summarized as the following.

- We propose a camera-independent learning color constancy method based on scene semantics (CISS). Compared with most learning-based and quasi-camera-independent methods, CISS generalizes well to unseen camera sensors without fine-tuning or re-training with additional images.
- We prove that the scene statistics related to gray-based assumptions are independent of the camera sensor. In addition, we also observed the correlation between these scene statistics and scenes.
- We provide an effective way to extract robust scene semantics without errors brought by camera-specific illuminant through the designed illuminant-invariance cross-moments.
- We evaluate CISS on the NUS and the INTEL-TAU datasets. On both datasets, CISS achieves superior performances compared to previous camera-independent methods for multi-cameras color constancy. In addition, CISS also outperforms existing color constancy methods based on image content understanding in the Shi-Gehler dataset.

2. Related Work

In this section, we first briefly discuss two color constancy works related to CISS: the camera-independent learning-based methods and image content understanding based methods. Then, we also introduce the gray-based assumptions used in CISS and show the flaws of these assumptions. Meanwhile, we account for the camera-independent property of the scene statistics related to gray-based assumptions.

2.1. Camera-independent learning-based methods

Recently, some camera-independent learning-based methods have been proposed to improve the generalization of models on datasets captured by different cameras. Hernandez et al. [15] propose to learn the likelihood of properly white-balanced images according to Bayesian formulation. Despite promising results, their methods require camera-specific candidate illuminants. Other strategies have formulated the multi-cameras color constancy as a few-shot learning [16] or multi-domain learning problem [17]. Both of them attempt to enable learning-based models to cope with differences between images taken by various cameras. Although the above methods effectively reduce the effort of re-training models, they still require fine-tuning on a set of test camera-specific additional images. The recently proposed C5 [13] reduces the constraint for these additional images, for example, these images can be unlabeled and not white balanced. However, selected additional images still require to be taken by the test camera. We argue that the need for additional camera-specific images weakens the generalizability of the trained model. Without these additional images, the above methods remain camera-specific. Therefore, we refer to them as quasi-camera-independent methods. Another class of work seeks to learn camera-independent color constancy models, circumventing the need to fine-tune the model according to additional images. A recent method [18] finds achromatic pixels with a CNN to predict the illuminant. Since in most camera sensors, achromatic pixels are rendered gray in the linear-RGB image under a neutral illuminant [2]. Another strategy [19] proposes to learn a camera-independent working space that can normalize the RGB values of any camera. The proposed model allows images captured from different cameras and therefore obtains competitive results in the multi-cameras task.

2.2. Image content understanding for color constancy

Another work related to CISS aims to solve color constancy problems through image content understanding. We name these methods as “Content Driven Methods” (CD). Early CD methods [14] aim to improve illuminant estimation accuracy by selecting or combining existing color constancy methods through image content understanding. More recently, some CD methods have started to obtain illuminant estimation cues from high-level visual information directly [20]. In addition, Bianco et al. [21] pre-trained the CNN with an image classification task. Then features extracted by the pre-trained CNN were used as input to linear support vector regression to estimate illuminant colors.

Similar to the recent CD method, CISS also utilizes high-level visual information of the image (scene semantics) to address the color constancy problem. The difference, however, is that CISS also overcomes the problem of poor model generalization when facing camera variations. Therefore, CISS performs well in multi-cameras color constancy tasks as well.

2.3. Gray-based assumptions

In color constancy, a type of well-established assumption is gray-based assumption: the scene statistics M_c^* in varying scenes under a neutral illuminant are achromatic [14]

$$M_c^* = \left(\iint \left| \nabla^n I_{c,\sigma}^*(x,y) \right|^p dxdy \right)^{1/p} = k \tag{3}$$

where $c = \{R,G,B\}$, n is the order of the derivative, k is a constant and p is the Minkowski-norm. $I_{c,\sigma}^*(x,y)$ is the convolution of the image under a neutral illuminant with a Gaussian filter G_σ with scale parameter σ. A wide variety of gray-based methods can be generated using Eq. 3. Based on the gray-based assumptions, any deviation from achromaticity in the scene statistics is caused by the effects of the illuminant. This implies that the illuminant E_c can be indirectly estimated by calculating the scene statistics of the image.

$$E_c = M_c / M_c^* \tag{4}$$

where M_c represents the scene statistics calculated from the color-biased image.

Gray-based methods are lightweight and comprehensible, and most of all, the assumptions of these methods are camera-

independent. Suppose that $K(\boldsymbol{I}_a)$ and $K(\boldsymbol{I}_b)$ represent the scene statistics of two white-balanced images captured by different cameras in the same scene under a neutral illuminant, respectively. The transforming from $K(\boldsymbol{I}_a)$ to $K(\boldsymbol{I}_b)$ can be formulated as [22].

$$\boldsymbol{T}_s K(\boldsymbol{I}_a) = K(\boldsymbol{I}_b) \tag{5}$$

where $\boldsymbol{T}_s$ is a 3×3 matrix that performs the transformation of color space between two different cameras. Since color differences caused by camera variations have been removed from the gray pixel in white-balanced images, the gray pixel is not disturbed by $\boldsymbol{T}_s$. Fortunately, based on gray-based assumptions, $K(\boldsymbol{I}_a)$ and $K(\boldsymbol{I}_b)$ are not supposed to deviate significantly from gray. As a result, matrix $\boldsymbol{T}_s$ ought not have a high impact on $K(\boldsymbol{I}_a)$ and $K(\boldsymbol{I}_b)$. This explains why the scene statistics distributions of images from different cameras are so consistent in Fig. 1 (b). Whereas, to a certain degree, even though gray-based assumptions are not disturbed by camera changes, these assumptions still cannot be universal in all images because of their fixed nature. Considering that the performance of gray-based methods has been proven to be affected by scene semantics [14], we selected four sets of scenes from [23] to analyze further the correlation between scene semantics and gray-based assumptions.

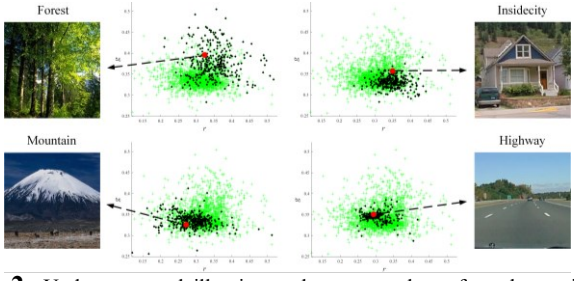


Fig. 2. Under a neutral illuminant, the scatter plots of *rg*-chromaticities distributions of the scene statistics related to gray world assumption in several scenes, overlayed on the *rg*-chromaticities distributions of the scene statistics from all scenes in the real-world set [23].

Fig. 2 shows the *rg*-chromaticities distributions of scene statistics related to the gray world assumption in four scenes. It can be seen that images with the same scene have similar *rg*-chromaticities distributions of the scene statistics. However, there are significant differences in the *rg*-chromaticities distributions for the scene statistics of images with different scenes. This shows the correlation between the scene semantics and the scene statistics related to gray-based assumptions, as well as intuitively explains why these fixed gray-based assumptions cannot be universal.

3. Method

Motivated by the correlation between scene semantics and the scene statistics related to gray-based assumptions, CISS estimates the scene statistic of each input image under a neutral illuminant based on scene semantics. The estimated scene statistic is further applied to the calculation of illuminant. Fig. 3 shows the flowchart of CISS.

3.1. Scene Semantic Calculation

3.1.1. Illuminant-Invariance Local Color Descriptors based on Cross-Moment

Feature extraction is a fundamental part of scene understanding. In this paper, we considered local color features since they are usually simple, fast, and rotationally invariant. Concretely, we extract local color features by calculating cross-moments from densely sampled image patches. These calculated cross-moments will later be utilized to form the image representation related to scene semantics. Unfortunately, in multi-cameras color constancy tasks, the illuminant of images taken by various cameras often differ significantly (see Fig. 1 (a)), which leads to instability of color features and reduces the accuracy of the formed image representation. To solve the above problem, we further improve cross-moment based on the Lambertian diffuse reflection model to enhance its robustness against illuminant variation. Based on the von Kries coefficient law, Eq. 1 can be given by the simplified diagonal model.

$$I_c(x,y) = E_c S_c(x,y) \tag{6}$$

where E_c represents the *c*-channel value of the illuminant $\boldsymbol{E}$ which is camera-specific, $S_c(x,y)$ represents the *c*-channel value of the image under a neutral illuminant [1]. Following [24], we define the generalized moment PM_{rq}^{abd} by

$$PM_{rq}^{abd} = \iint x^r y^q \left[I_R(x,y)\right]^a \left[I_G(x,y)\right]^b \left[I_B(x,y)\right]^d dxdy \tag{7}$$

where PM_{rq}^{abd} is the generalized moment of order and degree $a+b+d$. We only take into account the PM_{00}^{abd} in this paper because they are proven to be rotationally invariant. Then, we combine Eq. 6 and Eq. 7.

$$\begin{aligned} PM_{00}^{abd} &= \iint \left[I_R(x,y)\right]^a \left[I_G(x,y)\right]^b \left[I_B(x,y)\right]^d dxdy \\ &= \iint \left[E_R S_R(x,y)\right]^a \left[E_G S_G(x,y)\right]^b \left[E_G S_G(x,y)\right]^d dxdy \end{aligned} \tag{8}$$

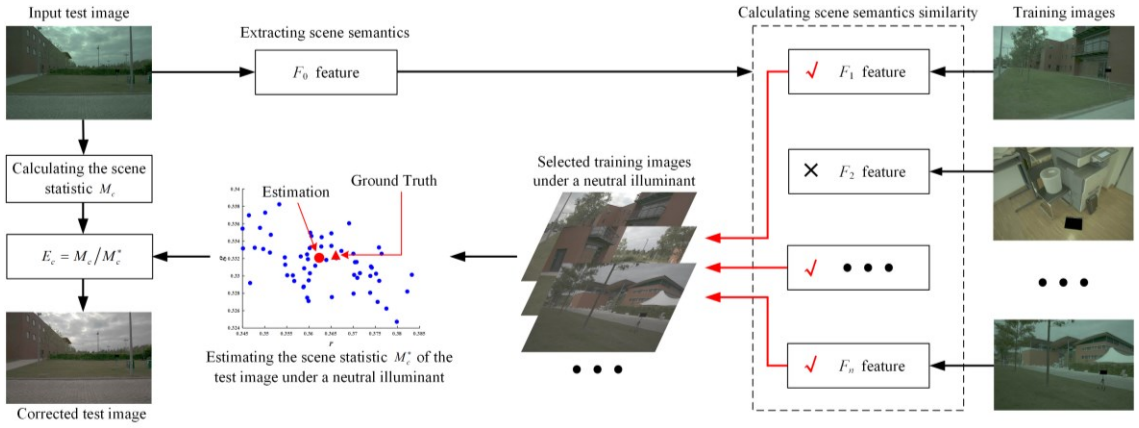


Fig. 3. The flowchart of CISS. For each input image, we extract the scene semantics with spatial pyramid structure using illuminant-invariance cross-moments as well as the combination of PLSA and spatial pyramid blocks. Then, we calculate the scene semantics similarity between the test image and each training image based on the pyramid match kernel, and select a set of color-corrected training images with higher similarity. These selected images will be further sifted for higher similarity through an automatic similarity threshold determined by the OTSU method. The remained images are used to estimate the scene statistic of the test image under a neutral illuminant. Finally, the estimated scene statistic is used to calculate the illuminant E_c for the test image.

Eq. 8 is the zero-order moment based on the image pixel values. We use the first-order derivative operator $f(\cdot)=\partial(\cdot)/\partial \boldsymbol{x}$ to extend Eq. 8 to edge moment further. For simplicity, we denote $f\left(I_c(x,y)\right)$ as $I_c^f(x,y)$ and $f\left(S_c(x,y)\right)$ as $S_c^f(x,y)$.

$$\begin{aligned} EM_{00}^{abd} &= \iint \left[I_R^f(x,y)\right]^a \left[I_G^f(x,y)\right]^b \left[I_B^f(x,y)\right]^d dxdy \\ &= \iint \left[E_R S_R^f(x,y)\right]^a \left[E_G S_G^f(x,y)\right]^b \left[E_G S_B^f(x,y)\right]^d dxdy \end{aligned} \tag{9}$$

Under the assumption that a single light source illuminates the scene, the illuminant E_c in Eq. 8 and Eq. 9 can be considered a constant and, therefore, able to be moved outside the integral. Based on this property, we can construct the following illuminant-invariance moment IM_{00}^{abd}.

$$IM_{00}^{abd} = PM_{00}^{abd} / EM_{00}^{abd} \tag{10}$$

In Eq. 10, The channel scaling of the camera-specific illuminant E_c has been removed, so that IM_{00}^{abd} is illuminant-invariance. Similar to [3], it is believed that cross-channel and higher degree provide useful color information. Therefore, the cross-moments up to the second degree with zero-order are considered in our work. The combination of these cross moments is denoted as (IM_{00}^{100}, IM_{00}^{010}, IM_{00}^{001}, IM_{00}^{200}, IM_{00}^{020}, IM_{00}^{002}, IM_{00}^{110}, IM_{00}^{101}, IM_{00}^{011})^T. The combination of the above cross-moments will be regarded as local color features IM of the image patch and used in the subsequent scene semantics extraction.

3.1.2. Extracting Scene Semantics with Spatial Pyramid Structure

Inspired by the success of the bag-of-words with spatial pyramid structure in scene semantic recognition, we perform the *K*-means method for all calculated local color features IM to form M visual words. Subsequently, these visual words are divided into different areas based on spatial pyramid blocks. Further, in order to obtain a more compact, discriminate representation, we treat all visual words in the same area as an independent document and further extract the probability distribution of topics in the document by PLSA.

Given a set of training documents $D=\{d_1,d_2,\ldots d_N\}$ with visual words from the visual word book $W=\{w_1,w_2,\ldots w_V\}$, where N and V denote the number of training documents and visual words. Furthermore, PLSA defines a set of latent variables (usually called 'topics') as $Z=\{z_1,z_2,\ldots z_R\}$, where R is the

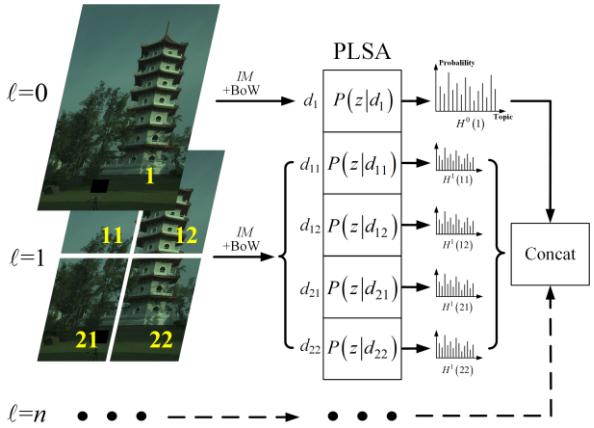


Fig. 4. The flowchart of extracting scene semantics with spatial pyramid structure

number of topics and is a hyperparameter. Assume that the visual word w is independent of the document d it belongs to, the joint probability over visual word and document $P(w,d)$ can be calculated as

$$P(w,d)=P(d)\sum_{z\in Z}P(w|z)P(z|d) \tag{11}$$

where $P(d)$ denotes the probability of observing d, $P(w|z)$ denotes the probability of observing w in topic-specific and $P(z|d)$ denotes the probability of observing z in document-specific. An efficient way to estimate the hidden variables in Eq. 11 is the Expectation-Maximization (EM) method, which alternates an expectation step for calculating posterior probabilities of latent variables z based on current estimates and a maximization step for updating parameters based on posterior probabilities until convergence. Once $P(z|d)$ is estimated, we treat $P(z|d)$ as a topic probability histogram $H^\ell(i)$ from a specific document, where ℓ and i account for the index of level and area in the spatial pyramid structure. Finally, we concatenate $H^\ell(i)$ from different levels and areas as the scene semantics of the whole image. An overview of the above procedure can be seen in Fig. 4.

3.2. Scene Semantic Similarity Calculation

As a measure for comparing the similarity of two histograms, the histogram intersection kernel (HIK) typically better performances than other commonly used measures, e.g., RBF kernel or l_2 distance [25]. Define H_X^ℓ and H_Y^ℓ as histograms from images X and Y under level ℓ, the similarity $I\left(H_X^\ell,H_Y^\ell\right)$ between H_X^ℓ and H_Y^ℓ based on HIK will be

$$I\left(H_X^\ell,H_Y^\ell\right)=\sum_{i=1}^{D}\min\left(H_X^\ell(i),H_Y^\ell(i)\right) \tag{12}$$

where $D=2^\ell$ and D is the total number of areas under level ℓ. $I\left(H_X^\ell,H_Y^\ell\right)$ is abbreviated as I^ℓ below. Further, the pyramid match kernel is used to combine the I^ℓ from different levels so as to calculate the scene semantic similarity between images X and Y.

$$\kappa^L(X,Y)=\frac{1}{2^L}I^0+\sum_{\ell=1}^{L}\frac{1}{2^{L-\ell+1}}I^\ell \tag{13}$$

3.3. Selecting Semantically Similar Images with OTSU

Based on the observation of Fig. 2, we contend that images sharing similar scene semantics tend to have similar scene statistics when they are under a neutral illuminant. With this assumption, we select similar training images for the test image by calculating the semantic similarity between the test image and each training image. The selected training images are then used to estimate the scene statistics of the test image. Specifically, We first fixedly select the top q similar images for each test image, which ensures that we can find a sufficient number of images in the training dataset for the test image. However, fixing the number of images to be selected is not a suitable strategy. This is because the number of similar images in the training set it is not fixed for different test images. Therefore, fixing the number of selected images in the training set will potentially introduce "noisy images" that are not similar to the test image. Therefore, we use OTSU to adaptively compute similarity thresholds for top q images selected in step one. Only images with a similarity score above the threshold are employed in the estimation of the scene statistics for the test image.

3.4. Illuminant Estimation

Once accurately selected images that are semantically similar to the test image, we estimate the approximate scene statistic M_c^* of the test image under a neutral illuminant by finding the median of the scene statistics from all selected images. Finally, the illuminant of the test image can be easily calculated by Eq. 4.

4. Evaluation

4.1. Datasets

We evaluate CISS using two multi-cameras color constancy datasets: the NUS 8-camera dataset [26] and the INTEL-TAU dataset [27]. The NUS dataset consists of 9 sub-datasets captured from different cameras, each containing approximately 210 images. To highlight the impact of the camera sensor, all sub-datasets comprise images representing the same scene with slight misalignments. Therefore, the NUS dataset is well suited for testing the impact of camera sensor on the color constancy method. Besides, the INTEL-TAU dataset contains a total of 7022 images from three sub-datasets captured by different cameras. Compared to the NUS dataset, the three sub-datasets of INTEL-TAU come from different scenes across 17 countries. Finally, we also compared the performance of CISS with the existing CD methods in Shi-Gehler dataset [28].

4.2. Implementing Details

4.2.1. Parameters Setting

In our experiment, we down-sample all images to 1080p. Subsequently, these images are uniformly split into 20×20 pixel size patches and used to calculate local color features. Then, we use 2000 visual words, 100 topics to extract scene semantics. We fixed the above parameters in our experiment since we found that these parameters slightly affected our experiment result. In particular, some optional parameters that are the level of the spatial pyramid L, the selected images' number q, and the parameters related to the gray-based methods (the gray-edge order n, the Minkowski-norm p, and the standard deviation σ of Gaussian filter G_σ) need to be set in advance. These parameters are selected out form $L\in\{1,2,3\}$, $q\in\{10,20,\ldots,100\}$, $n\in\{1,2\}$ and $p,\sigma\in\{1,3,6\}$ through 3-fold cross-validation on the training dataset in each experiment.

4.2.2. Performance Metric

We choose the angular error (AE), which is widely used in the evaluation of color constancy methods, as the error metric.

$$\varepsilon_{angle}=\arccos\left(\frac{\boldsymbol{E}_e\cdot\boldsymbol{E}_t}{\|\boldsymbol{E}_e\|\cdot\|\boldsymbol{E}_t\|}\right) \tag{14}$$

where $\boldsymbol{E}_e\cdot\boldsymbol{E}_t$ is the dot product of the estimated illuminant $\boldsymbol{E}_e$ and the ground truth $\boldsymbol{E}_t$. To summarize the performance of each method across all images, we provide five statistical metrics, including the mean, median and trimean of the AEs, the mean of the best 25% and the worst 25% AEs, as done in [1].

4.3. Experiment results

We evaluate the proposed CISS in both single-camera and multi-cameras scenarios using the NUS and INTEL-TAU datasets. In the single-camera setting, we follow most of the work performing 3-fold cross-validation on the dataset captured by the same camera, ensuring no camera sensor differences between the training and test images. Specifically, these images taken from the same camera are divided equally into three groups according to image number, and each time two sets are used for training while the remaining image set is used for testing. We also perform multi-cameras evaluation utilizing sub-datasets from the NUS or INTEL-TAU dataset, as both datasets contain sub-datasets captured by different cameras. Similar to *N*-fold cross-validation, one sub-datasets are used for testing each time while the remaining image set is used for training.

The comparison color constancy methods in our experiment are classified into two types, the learning-free methods and learning-based methods. All learning-free methods are camera-independent [2], and only a few learning-based methods share this property. It should be noted that our experimental results without quasi-camera-independent methods like [13, 15-17], which require additional specific-camera images or illuminants to fine-tune the model. It would lead to an unfair comparison for all learning class methods since all input images should be camera agnostic in the test phase. Instead, we provide two recent camera-independent methods, Quasi-Unsupervised [18] and SIIE [19]. Similar to CISS in this paper, neither requires any camera-specific additional information to adapt to the new camera.

4.3.1. NUS Dataset

Table 1 presents the accuracy of different methods on the NUS dataset in terms of several AEs statistical metrics. The performance of learning-free methods demonstrated only on the side of multi-cameras color constancy, since different cross-validation strategies make no difference to the performance of these methods. The best and second AEs statistical metrics is shown in red and blue, respectively.

Table 1. Quantitative evaluation of color constancy methods on the NUS dataset. All values correspond to AEs statistical metrics in degrees.

Methods		Multi-Cameras Color Constancy					Single-Camera Color Constancy				
		Mean	Median	Trimean	Best 25%	Worst 25%	Mean	Median	Trimean	Best 25%	Worst 25%
Learning-free	Gray-world [9]	4.22	3.21	3.46	0.92	9.27	-	-	-	-	-
	White-Patch [10]	12.26	14.53	13.02	1.88	21.59	-	-	-	-	-
	Shades-of-Gray [11]	4.22	3.21	3.46	0.92	9.27	-	-	-	-	-
	1st-order Gray-Edge [12]	3.71	2.81	3.06	0.85	7.95	-	-	-	-	-
	Cheng et al. 2014 [26]	4.30	2.95	3.31	0.79	10.04	-	-	-	-	-
	Gray Index [2]	3.18	2.22	2.42	0.61	7.44	-	-	-	-	-
Learning-based	CM (Pixel) [3]	2.80	2.35	2.46	0.87	5.52	2.73	1.89	2.06	0.60	6.33
	SIRMF (Pixel) [29]	2.63	2.10	2.23	0.77	5.34	2.49	1.84	1.97	0.57	5.55
	Cheng et al. 2015 [5]	2.59	1.99	2.12	0.63	5.61	2.42	1.61	1.78	0.50	5.76
	LCC [4]	2.67	1.91	2.11	0.74	5.81	2.47	1.53	1.76	0.50	5.99
	FFCC [6]	2.44	1.85	1.98	0.61	5.26	2.66	1.59	1.79	0.50	6.70
	SIIE [19]	2.05	1.50	-	0.52	4.48	-	-	-	-	-
	Quasi-Unsupervised [18]	1.97	1.41	-	-	-	4.00	2.86	3.17	0.86	8.95
	CISS (Proposed)	1.20	0.82	0.89	0.29	2.80	3.10	2.41	2.54	0.84	6.54

According to Table 1, LCC achieves the best overall performance in the single-camera setting. In fact, owing to inferring the valid information in prior training images, nearly all learning-based methods outperform learning-free methods in the single-camera setting. However, from Table 1, we also observed the vast majority of learning-based methods perform worse in the multi-cameras setting. For example, the average AE of regression-based methods (e.g., CM and SIRMF) increased by 24% and 14%, respectively, in the multi-camera setting compared to that in the single-camera setting. Since these methods typically precisely learn the mapping from image features to camera-specific illuminants, it is not surprising that

performance degrades in the camera agnostic scenario. It is worth mentioning that both SIRMF and CISS perform illuminant estimation by selecting suitable images. However, our method (CISS) differs from SIRMF in two points. First, SIRMF selects images with similar illuminants, while our method selects images with similar scenes. Second, SIRMF estimates camera-specific illuminant directly by a correction matrix, and therefore SIRMF is still affected by camera variations. In contrast, CISS is camera-independent. Interestingly, we found learning-free methods generally outperform learning-based methods in the multi-cameras setting. The reason being that learning-free methods estimate illuminant independently on a per-image basis, thus not significantly affected by camera variations [2]. Nevertheless, consider that the information from training images is not available. These methods depend on fixed assumptions usually, thus their performance is limited. In contrast, the camera-independent learning-based methods enable efficiently leverage prior information from training images without disturbing camera variations. As shown in Table 1, the Quasi-Unsupervised, SIIE, and CISS (the proposed) show excellent performance in the multi-cameras setting. Particularly, CISS achieves the best performance among all compared methods, and the AEs statistical metrics in the multi-cameras setting are noticeably reduced compared within the single-camera setting. This result proves that CISS can benefit from more training images without taking into account the possible differences of cameras when these images are captured.

4.3.2 INTEL-TAU Dataset

Table 2 lists the AEs statistical metrics of various color constancy methods on INTEL-TAU dataset.

Table 2. Quantitative evaluation of color constancy methods on the INTEL-TAU dataset.

Methods		Multi-Cameras Color Constancy					Single-Camera Color Constancy				
		Mean	Median	Trimean	Best 25%	Worst 25%	Mean	Median	Trimean	Best 25%	Worst 25%
Learning-free	Gray-world [9]	4.93	3.87	4.14	0.96	12.70	-	-	-	-	-
	White-Patch [10]	7.0	5.4	6.2	1.1	14.6	-	-	-	-	-
	Shades-of-Gray [11]	4.0	2.9	3.2	0.7	9.0	-	-	-	-	-
	1st-order Gray-Edge [12]	5.3	4.1	4.5	1.0	11.7	-	-	-	-	-
	Cheng et al. 2014 [26]	4.43	3.03	3.40	0.71	10.46	-	-	-	-	-
	Gray Index [2]	3.86	2.32	2.75	0.51	9.74	-	-	-	-	-
Learning-based	CM (Pixel) [3]	4.17	3.68	3.76	1.52	7.68	3.10	2.26	2.46	0.64	7.01
	SIRMF (Pixel) [29]	3.27	2.45	2.66	0.78	7.14	3.02	2.15	2.35	0.58	6.95
	Cheng et al. 2015 [5]	5.21	4.68	4.73	1.83	9.58	3.26	2.01	2.34	0.51	8.11
	LCC [4]	3.76	3.05	3.20	0.86	7.92	2.69	1.63	1.84	0.42	6.79
	FFCC [6]	3.21	2.43	2.55	0.67	7.18	2.59	1.57	1.77	0.41	6.55
	SIIE [19]	3.42	2.42	2.64	0.73	7.80	-	-	-	-	-
	Quasi-Unsupervised [18]	3.25	2.15	2.43	0.60	7.74	3.55	2.42	2.70	0.65	8.34
	CISS (Proposed)	2.74	1.95	2.12	0.54	6.32	2.97	2.19	2.36	0.66	6.63

Similar to Table 1, we can observe in Table 2 that mostly learning-based methods with excellent performance in the single-camera setting still have a notable degradation in the multi-cameras setting. Besides, CISS has once again achieved optimal performance in terms of various AEs statistical metrics in the multi-cameras setting. It is worth noting that the performance improvement of CISS on the INTEL-TAU dataset with multiple camera settings is not significant compared to the NUS dataset. We argue that it could be caused by the low correlation of scenes in different sub-datasets, as the scenes in each sub-dataset of the INTEL-TAU dataset come from different countries worldwide. Nonetheless, CISS still obtains a slight performance gain in the multi-cameras setting without a significant performance drop like most learning class methods.

4.3.3. Ablation Study and Analysis

In the following ablation study, we further investigate the contribution of each component of CISS. We establish the following four baseline methods. Table 3 shows the results of CISS and these baseline methods in the ablation study.

1) CISS (PM) and CISS (EM): the illuminant-invariance property of the local color descriptor is removed and instead, we use PM_{00}^{abd} and EM_{00}^{abd} (see Eq. 8 and Eq. 9) to extract local color features.

2) CISS (w/o SP): the spatial pyramid structure is removed from scene semantics.

3) CISS (w/o OTSU): estimating the scene statistics of the test image only using the top q similar images fixedly.

As shown in Table 3, the CISS performance drops dramatically in both datasets when the cross-moments PM_{00}^{abd} and EM_{00}^{abd} are used. We argue that the possible reason for the above problem is that the robustness of the scene semantics is

Table 3. Results of ablation studies on the NUS and INTEL-TAU datasets.

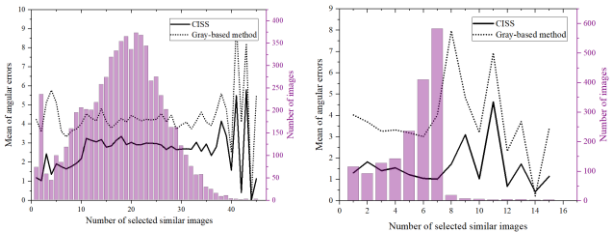
Methods	NUS			INTEL-TAU		
	Mean	Median	Trimean	Mean	Median	Trimean
CISS (PM)	1.48	0.95	1.06	3.61	2.94	3.07
CISS (EM)	1.39	0.92	1.03	3.75	3.17	3.28
CISS (w/o SP)	1.36	0.88	0.96	2.91	2.20	2.37
CISS (w/o OTSU)	1.32	0.85	0.94	2.97	2.22	2.39
CISS	1.20	0.82	0.89	2.74	1.95	2.12

degraded due the PM_{00}^{abd} and EM_{00}^{abd} are unable to remove camera-specific illuminant, which affects the accuracy of inferred gray-based assumptions and eventually leads to CISS performance degradation. In addition, it is clear that removing the spatial pyramid structure in scene semantics also drops the CISS performance, which means the spatial pyramid structure contributes to the accurate extraction of scene semantics and further helps CISS reach a better performance. In summary, the CISS of combining all components is always superior to other baseline methods under most AEs statistical metrics. These results validates the effectiveness of the proposed components.

4.4. Analysis of Image Selection Quantity and Illumination Estimation Error

Based on the image selection strategy described in Section 3.3, CISS estimates the scene statistics of the test image by selecting a set of similar images and indirectly computes the illumination of the test image. To show the robustness of CISS in estimating the illumination under different numbers of selected images, we first divided the test images based on the number of training images selected when estimating the scene statistics. In each subfigure of Fig. 5, the X-axis represents the number of training images selected for the test image, while the bar chart and the left Y-axis represent the total number of test images selected with the

current number of training images. We also plotted a line graph on the left *Y*-axis to show the average angular error of illuminant estimation for test images when using the number of training images displayed on the *X*-axis. The dashed line chart above represents the angular error when no image selection is performed, i.e., using all training images to estimate scene statistics and compute illumination, which represents the traditional gray-based method.



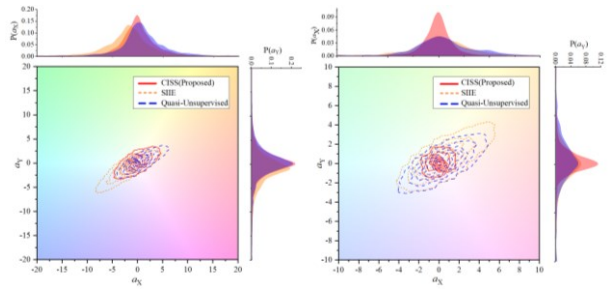
(a) INTEL-TAU (b) NUS

Fig. 5. Relationship between image selection number and illumination estimation error on two datasets

It should be noted that only a few test images select a large number of images, so the mean of the angular error calculated based on these test images is not stable. Therefore, we recommend ignoring the sharp fluctuations at the end of the line. Observing Figure 5, it can be seen that CISS achieves stable and superior performance in light source estimation compared to the traditional gray-based method, even when selecting a small number of training images. This indicates that CISS is robust to the number of selected training images.

4.4. Visualization error analysis

In this section, we visualize the illuminant estimation error of CISS with two other camera-independent methods SIIE with Quasi-Unsupervised using Angle-Retaining Chromaticity diagram [30].



(a) (b)

Fig. 6. Error distributions for CISS, SIIE and Quasi-Unsupervised on (a) the INTEL-TAU dataset and (b) the NUS dataset. a_X and a_Y correspond to the distance in the ARC Cartesian coordinate between the perfect white surface with the reproduction of a white surface (corresponding to the ground truth illuminant) corrected using the estimated illuminant. $P(a_X)$ and $P(a_Y)$ represent the probability distributions of a_X or a_Y on the corresponding axes.

In Fig. 6, The closer a_X and a_Y are to 0 (the diagram center), the better illuminant estimation performance. It can be seen that in the INTEL-TAU dataset, CISS closer to the diagram center. In addition, the other methods display a different hue-specific bias, with Quasi-Unsupervised being more spread toward the magenta region of the diagram, and SIIE toward the opposite end. In addition, CISS exhibits a more isotropic distribution in the NUS dataset, compared to the skewed results of SIIE and Quasi-Unsupervised.

4.5. Compared to CD methods

In spite of our method's focus on the problem of poor model generalization when facing camera variations. However, CISS also belongs to CD methods because of the use of scene semantic. We compare the illuminant estimation performance of CISS with existing CD methods using the Shi-Gehler dataset in Table 4.

Table 4. Quantitative evaluation of CD methods on the Shi-Gehler dataset.

Methods	Mean	Median	Trimean
Natural Image Statistics [14]	**4.40**	**3.18**	**3.49**
Multi-cue-based methods [20]	**3.25**	**2.20**	**2.55**
Exemplar-Based [31]	**4.40**	**3.30**	**-**
AlexNet+SVR [21]	**4.74**	**3.09**	**-**
Buzzelli et al. 2018 [32]	**4.84**	**4.12**	**-**
CISS (Proposed)	**2.69**	**1.95**	**2.10**

As shown in Table 4, CISS outperforms existing CD methods in all illuminant estimation performance metrics. Thus, besides proposing a camera-independent learning-based method, another contribution of our paper is to reveal further the potential of image content understanding in illuminant estimation.

4.6. Visual Comparison

Fig. 7 shows a visual comparison of the results of the proposed CISS with other methods. Observing the selected images of CISS in the fifth column of Fig. 6, we can see that CISS always returns similar images for most of the input image. Most importantly, the performance of CISS will not be degraded by different camera models of these images, and this camera-independent property is typically not shared by most learning-based methods. Besides, it is interesting to note that CISS provides more acceptable results than gray-based methods in most images (e.g., the image in lines 2, 3 and 4). It is due to the gray-based assumptions considering the scene statistics as a constant value. In contrast, CISS is able to estimate the scene statistics of the test image accurately by scene semantics and then calculate the illuminant using the estimated scene statistics. In addition, it can be observed the Quasi-Unsupervised method seems to have a poor performance in certain scenes (e.g., the image in line 3), which could be caused by the absence of recognizable gray pixels in these scenes. Finally, it is noteworthy that the CISS exhibits poor performance in line 4 image. This reason is mainly attributed to the fact that the scene semantics of the test image is scarce, making it difficult for CISS to accurately estimate the scene statistics of the test image. Therefore, our future work will focus on semantic data augmentation to improve the performance of color constancy methods.

5. Conclusion and Discussion

We have presented CISS, a camera-independent learning method based on scene semantics. Inspired by the camera-independent property of the gray-based methods. CISS estimates camera-independent scene statistics related to gray-based assumptions by scene semantics and performs illuminant estimation, thus avoiding the impact of camera variations on the model. Experiments indicate that CISS outperforms all gray-based methods and existing camera-independent methods for multi-cameras color constancy on several datasets, while being capable of applying to camera-agnostic images without fine-tuning with additional images.

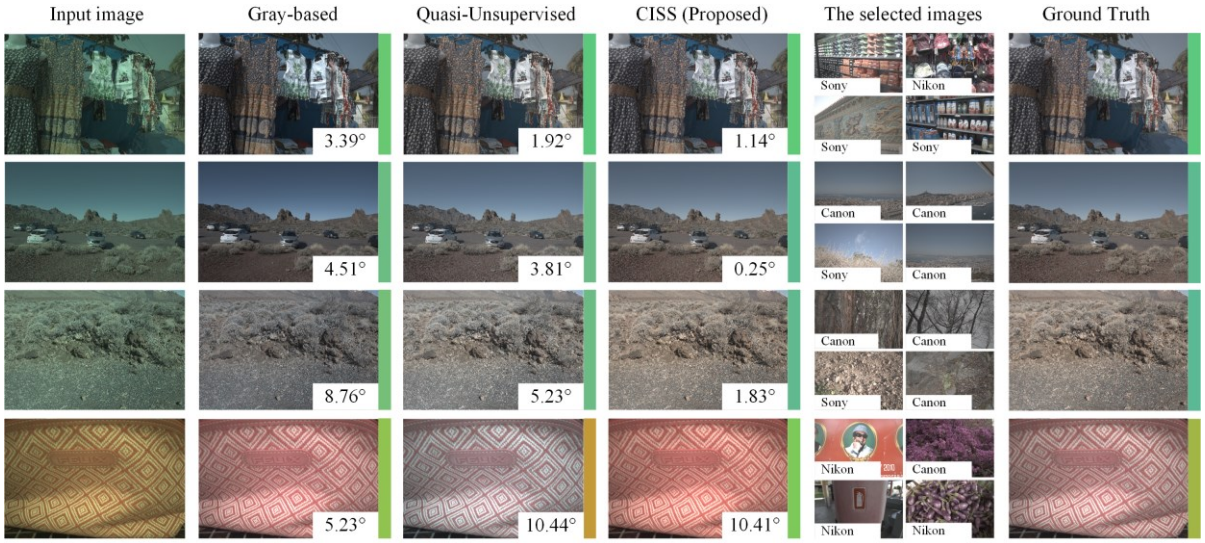


Fig. 7. Corrected images from three methods including proposed CISS. The AE is marked at the lower-right corner of the corrected image. The global illuminant for image correction is given on the right side of the images. In addition, the fifth column shows a part of images selected by CISS that have similar scenes with the input image. The camera model of these selected images is given at the lower-left corner of the image.

Acknowledgments

This work was supported by the National Key Research and Development Program of China [2017YFC0822204]; the Technical Research Program of Ministry of Public Security [2020JSYJC25]; Open Project of Key Laboratory of Forensic Science of Ministry of Justice [KF202014]; Innovative Talents Support Program of Liaoning Province [LNCX202007] and the Young scientific and technological talents breeding project [JYT2020130] and the National Natural Science Foundation of China under Grants [NO. 62072126].

References

1. Gijsenij, A., Gevers, T., van de Weijer, J., 2011. Computational Color Constancy: Survey and Experiments. IEEE Trans. Image Process 20, 2475-2489.
2. Qian, Y., Kamarainen, J.-K., Nikkanen, J., et al., 2019. On Finding Gray Pixels, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 8054-8062.
3. Finlayson, G.D., 2013. Corrected-Moment Illuminant Estimation, Proc. IEEE Int. Conf. Comput. Vis., pp. 1904-1911.
4. Chakrabarti, A., 2015. Color Constancy by Learning to Predict Chromaticity from Luminance, Proc. Advances in Neural Information Processing Systems.
5. Cheng, D., Price, B., Cohen, S., et al., 2015. Effective Learning-Based Illuminant Estimation Using Simple Features, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 1000-1008.
6. Barron, J.T., Tsai, Y.-T., 2017. Fast Fourier Color Constancy, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 6950-6958.
7. Lo, Y.-C., Chang, C.-C., Chiu, H.-C., et al., 2021. CLCC: Contrastive Learning for Color Constancy, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 8053-8063.
8. Xu, B., Liu, J., Hou, X., et al., 2020. End-to-End Illuminant Estimation based on Deep Metric Learning, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 3613-3622.
9. Buchsbaum, G., 1980. A spatial processor model for object colour perception. J. Frankl. Inst. 310, 1-26.
10. Land, E.H., 1977. The retinex theory of color vision. Scientific American 237, 108-128.
11. Finlayson, G.D., Trezzi, E., 2004. Shades of gray and colour constancy, Proc. IS&T/SID 12th Color Imaging Conf., pp. 37-41.
12. Van de Weijer, J., Gevers, T., Gijsenij, A., 2007. Edge-based color constancy. IEEE Trans. Image Process 16, 2207-2214.
13. Afifi, M., Barron, J.T., LeGendre, C., et al., 2021. Cross-camera convolutional color constancy, Proc. IEEE Int. Conf. Comput. Vis., pp. 1981-1990.
14. Gijsenij, A., Gevers, T., 2011. Color Constancy Using Natural Image Statistics and Scene Semantics. IEEE Trans. Pattern Anal. Mach. Intell. 33, 687-698.
15. Hernandez-Juarez, D., Parisot, S., Busam, B., et al., 2020. A Multi-Hypothesis Approach to Color Constancy, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 2267-2277.
16. McDonagh, S., Parisot, S., Zhou, F., et al., 2018. Formulating Camera-Adaptive Color Constancy as a Few-shot Meta-Learning Problem. arXiv preprint arXiv:1811.11788.
17. Xiao, J., Gu, S., Zhang, L., 2020. Multi-Domain Learning for Accurate and Few-Shot Color Constancy, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 3255-3264.
18. Bianco, S., Cusano, C., 2019. Quasi-Unsupervised Color Constancy, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 12204-12213.
19. Afifi, M., Brown, M.S., 2019. Sensor-independent illumination estimation for DNN models. arXiv preprint arXiv:1912.06888.
20. Li B, Xiong W, Hu W, et al., 2016. Multi-cue illumination estimation via a tree-structured group joint sparse representation. Int. J. Comput. Vis. 117: 21-47.
21. Bianco, S., Cusano, C., Schettini, R., 2017. Single and multiple illuminant estimation using convolutional neural networks. IEEE Trans. Image Process 26, 4347-4362.
22. Nguyen, R., Prasad, D.K., Brown, M.S., 2014. Raw-to-raw: Mapping between image sensor color responses, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 3398-3405.
23. Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 2169-2178.
24. Wu, M., Luo, K., Dang, J.J., et al., 2015. Edge-moment-based color constancy using illumination-coherent regularized regression. J. Opt. Soc. Am. A. 32, 1707-1716.
25. Wu, J.X., 2010. A Fast Dual Method for HIK SVM Learning, Proc. Eur. Conf. Comput. Vis., pp. 552-565.
26. Cheng, D., Prasad, D.K., Brown, M.S., 2014. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. J. Opt. Soc. Am. A. 31, 1049-1058.
27. Laakom, F., Raitoharju, J., Nikkanen, J., et al., 2021. INTEL-TAU: A Color Constancy Dataset. IEEE Access 9, 39560-39567.
28. Gehler, P.V., Rother, C., Blake, A., et al., 2008. Bayesian color constancy revisited, Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 1-8.
29. Gao, S. B., Zhang, M., Li, Y. J., 2019. Improving color constancy by selecting suitable set of training images. Opt. Express 27(18), 25611-25633.
30. Buzzelli, M., Bianco, S., Schettini, R., 2020. ARC: Angle-Retaining Chromaticity diagram for color constancy error analysis. J. Opt. Soc. Am. A. 37, 1721-1730.
31. Joze, H.R.V., Drew, M.S., 2014. Exemplar-Based Color Constancy and Multiple Illumination. IEEE Trans. Pattern Anal. Mach. Intell. 36, 860-873.
32. Buzzelli, M., van de Weijer, J., Schettini, R., 2018. Learning illuminant estimation from object recognition, Proc. IEEE Int. Conf. Image Process., pp. 3234-3238.