Multi-Cue Semi-Supervised Color Constancy With Limited Training Samples

Xinwei Huang, Bing Li, Shuai Li, *Member, IEEE*, Wenjuan Li, Weihua Xiong, Xuanwu Yin, Weiming Hu, *Senior Member, IEEE*, and Hong Qin

***Abstract*— Color constancy is one of the fundamental tasks in computer vision. Many supervised methods, including recently proposed Convolutional Neural Networks (CNN)-based methods, have been proved to work well on this problem, but they often require a sufficient number of labeled data. However, it is expensive and time-consuming to collect a large number of labeled training images with accurately measured illumination. In order to reduce the dependence on labeled images and leverage unlabeled ones without measured illumination, we propose a novel semi-supervised framework with limited training samples for illumination estimation. Our key insight is that the images with similar features from different cues will share similar lighting conditions. Consequently, three graphs based on three visual cues, low-level RGB color distribution, mid-level initial illuminant estimates and high-level scene content, are constructed to represent the relationship among different images. Then a multi-cue semi-supervised color constancy method (MSCC) is proposed after integrating these three graphs into a unified model. Extensive experiments on benchmark datasets demonstrate that our proposed MSCC method outperforms nearly all the existing supervised methods with limited labeled samples. Even with no unlabeled samples, MSCC still obtains better performance and stableness than most supervised methods.**

***Index Terms*— Color constancy, illumination estimation, white balancing, multi-cue, semi-supervised.**

Manuscript received August 31, 2019; revised January 27, 2020 and May 16, 2020; accepted June 29, 2020. Date of publication July 13, 2020; date of current version July 22, 2020. This work was supported in part by the Beijing Natural Science Foundation under Grant JQ18018 and Grant L172051, in part by the National Key Research and Development Program of China under Grant 2017YFB1002801, Grant 2016QY01W0106, and Grant 2017YFF0106407, in part by the National Natural Science Foundation of China under Grant U1936204, Grant U1803119, Grant U1736106, Grant 61532002, Grant 61906192, and Grant 61876100, and in part by the USA NSF under Grant IIS-1715985 and Grant 1812606. The work of Bing Li was supported in part by the Youth Innovation Promotion Association, CAS. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiantao Zhou. *(Corresponding author: Bing Li.)*

Xinwei Huang and Shuai Li are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China.

Bing Li, Wenjuan Li, and Weihua Xiong are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Beijing 100190, China, and also with PeopleAI Inc., Beijing 100190, China (e-mail: bli@nlpr.ia.ac.cn).

Xuanwu Yin is with the Department of Kirin Chipset and Technology Development, Hisilicon, Beijing 100095, China.

Weiming Hu is with the CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

Hong Qin is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA.

Digital Object Identifier 10.1109/TIP.2020.3007823

I. Introduction

COLOR constancy, which aims at correcting image's color deviations caused by a difference in illumination as done by the human vision system [1], [2], has become an important problem in several important applications in the field of computer vision, such as auto white balancing, object recognition, image matching and visual tracking. It normally includes two steps: obtaining an estimate of the light color and computing illumination independent surface descriptor under the help of Von Kries Diagonal transformation [3]. Therefore, the illumination estimation is the key to the color constancy.

Many methods have been proposed to solve the illumination estimation problem. Supervised methods, especially recently proposed CNN-based methods, have achieved leading performance with large scale datasets. However, labeling a large training dataset with accurate illumination values is usually done using human expertise, which is expensive, time-consuming and error-prone. A standard object with known chromatic properties, such as a gray ball or a color checker, is usually required when taking images for the datasets used for color constancy problem. The object is used to calculate the ground-truth illumination of the image. It is clear that the procedure is impractical. In fact, the benchmark datasets only contain up to hundreds or thousands of images, which are much less than those used for other computer vision tasks. Besides, most supervised methods need to train different models for different types of cameras, which means that one needs to strenuously collect and label a new dataset for every new camera type. Conversely, obtaining unlabeled data is usually much easier since it only involves collecting images without having to give out their ground-truth illumination. For example, it is easy to collect lots of images with a specific camera, but it is very tedious and difficult to get the exact illumination value for each image.

In this paper, leveraging both labeled and unlabeled samples, we propose a graph-based semi-supervised color constancy algorithm (SCC), in which the relationship among samples (including both labeled and unlabeled ones) is represented as a graph, and then a semi-supervised regression model is trained on the graph for illumination estimation. We construct the graph using different features from different visual cues, respectively: low-level color histogram (LL), mid-level initial illumination estimates (ML) and high-level scene category (HL). Furthermore, in order to improve the stability and

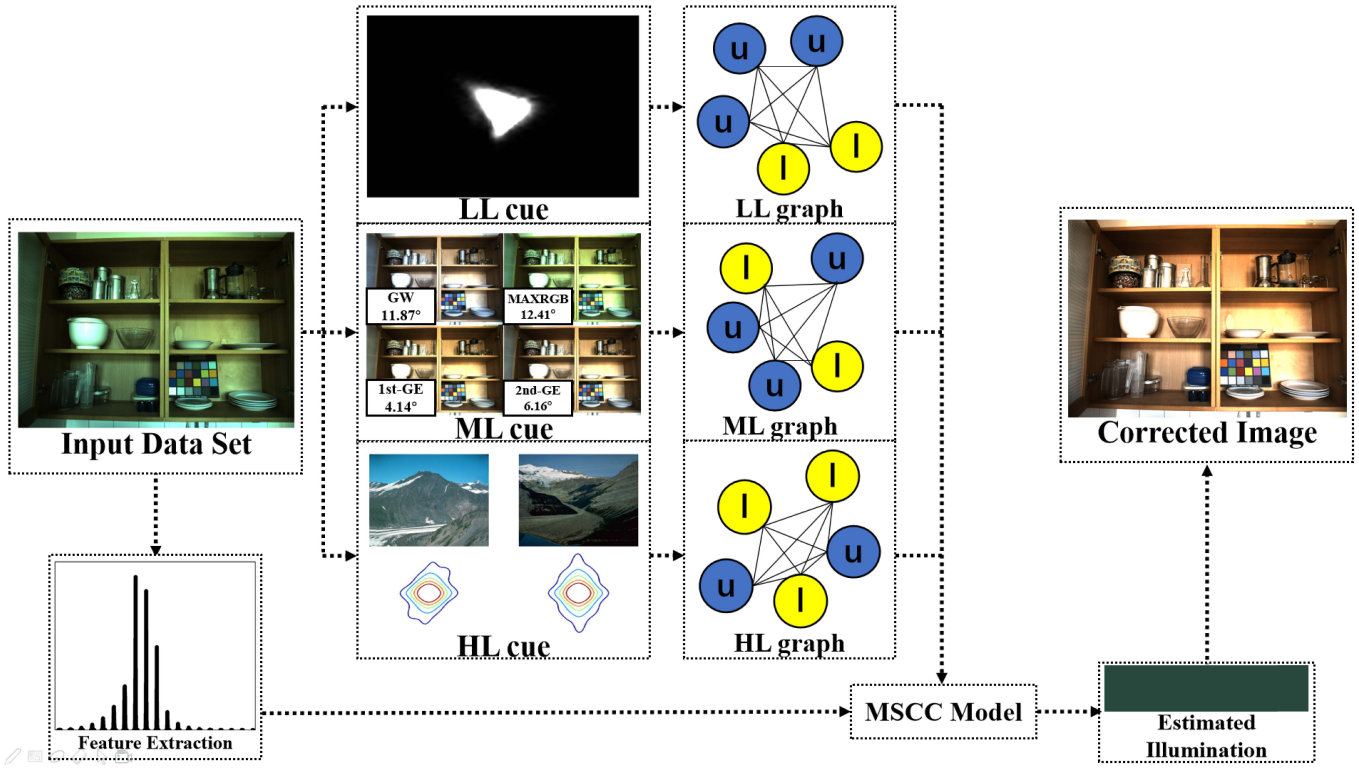


Fig. 1. An overview of our proposed multi-cue semi-supervised color constancy. Given a dataset including both labeled and unlabeled samples, we construct the relationship among them using graphs from three visual cues: low-level color histogram (LL), mid-level initial illumination estimates (ML) and high-level scene category (HL). Then three cues are integrated into a unified semi-supervised regression model for illumination estimation.

performance of the SCC, we integrate these three cues into a unified semi-supervised framework called multi-cue semi-supervised color constancy algorithm (MSCC). An overview of the proposed MSCC is shown in Fig.1.

The contributions of this paper can be summarized as follows:

- It proposes a semi-supervised color constancy (SCC) method to solve the problem of insufficient labeled training samples. Three different cues are introduced to construct the graph in SCC. As far as we are aware, this paper is the first to use the semi-supervised method in the color constancy field.
- It extends the SCC to a multi-cue semi-supervised color constancy method (MSCC) by integrating low-level, mid-level, and high-level visual cues into a unified framework to improve stability and performance of illumination estimation.
- Experimental results on different datasets show that the proposed MSCC not only outperforms nearly all the supervised and unsupervised methods with limited training samples, but also achieves comparable performance to state-of-the-art supervised methods without unlabeled samples.
- Experimental results show that the proposed methods can achieve better results if more unlabeled data are added. This is an important advantage for practical color constancy.

The remainder of this paper is organized as follows: Section II introduces the image model and related work. Section III presents the details of the graph-based semi-supervised color constancy. Section IV extends the SCC into the MSCC. Section V demonstrates the experimental results. Section VI concludes the paper.

II. Background

A. Imaging Model

The color signal $\mathbf{I}(\varepsilon) = [I_R(\varepsilon), I_G(\varepsilon), I_B(\varepsilon)]^T$ recorded by a camera for light reflected from a matte surface at spatial coordinate ε depends on three factors: the surface reflectance, $\mathbf{q}(\varepsilon, \lambda)$, the spectral power distribution of the incident light, $\mathbf{p}(\lambda)$, and the camera's spectral sensitivity functions, $\rho(\lambda) = [\rho_R(\lambda), \rho_G(\lambda), \rho_B(\lambda)]^T$:

$$\mathbf{I}_\theta(\varepsilon) = \int_\omega \mathbf{p}(\lambda)\mathbf{q}(\varepsilon, \lambda)\rho_\theta(\lambda)d\lambda, \quad \theta = \{R, G, B\}, \tag{1}$$

where λ indicates wavelength, and Ω is the visible spectrum interval. For a specific scene, we assume that the relative spectral power distribution remains the same. For the case of an ideal 'white' reflectance, we obtain the color signal corresponding to the illumination $\mathbf{e}(\varepsilon) = [e_R(\varepsilon), e_G(\varepsilon), e_B(\varepsilon)]^T$ as:

$$e_\theta(\varepsilon) = \int_\Omega \mathbf{p}(\lambda)\rho_\theta(\lambda)d\lambda, \quad \theta = \{R, G, B\}. \tag{2}$$

Thus, illumination estimation from an image is an ill-posed and under-determined inverse problem due to collinearity between object color and illuminant color.

B. Related Work

Illumination estimation has been the subject of a large body of research and many different methods have been proposed

in both scientific community and imaging industry for several decades. Most of these methods generally rely on some kinds of assumptions and can be roughly divided into two major categories: unsupervised methods and supervised methods.

1) Unsupervised Methods: The methods belonging to the unsupervised category explicitly predefine illumination estimation models based on certain hypotheses. MaxRGB [4] algorithm estimates the illumination based on the maximum response found within different color channels, and then is improved with some simple preprocessing operations by Funt and Shi [5]. Grey World (GW) algorithm proposed by Buchsbaum [6] assumes that the average of the channels taken separately represents the illuminant color signal. The MaxRGB and GW are then generalized to Shades of Grey algorithm (SoG) [7] using Minkowski-norm. Grey Edge (GE) [8] method assumes that the average reflectance differences in a scene are achromatic. Furthermore, a Grey Edge framework is introduced to unify a variety of unsupervised methods by including higher-order and derivatives. By concerning the role of the retinal mechanism of the biological visual system, Zhang *et al.* [9] propose a novel color constancy method that models the double-opponent (DO) cells of the human visual system. Cheng and Brown [10] assume that large color differences are the key to estimating the illumination. Intuitively, bright and dark pixels are chosen using a projection distance for illumination estimation. Yang *et al.* [11] hypothesize that there are grey pixels widely appearing in natural scenes. By calculating a Grey Index (GI) for each pixel, those grey pixels are detected and used for illumination estimation. Recently, Bianco and Cusano [12] propose a quasi-unsupervised method, in which a pre-trained deep convolutional neural network is exploited to obtain a weighted average of detected achromatic pixels. Although the method does not require illumination information, a training phase is still necessary.

In general, unsupervised methods are simple and have much lower complexity. However, the fixed estimation models embedded in them result in lower generalization. Once the model is selected, the illumination colors of all the test images are computed out using the same model. Therefore, the methods are effective only when the distribution of colors of the test image fits the assumed model very well. Besides, existing unsupervised methods mainly concentrate on low-level statistical information and ignore useful high-level semantic information, which leads to an unsatisfactory performance compared to supervised ones.

2) Supervised Methods: Supervised methods learn the estimation models on the color distribution or related features of training data. Color by correlation (CbyC) [13] constructs a correlation matrix which describes the interrelation between illumination values and image chromaticity distributions. Then the illumination with the highest probability is chosen. CbyC is extended by Vazquez-Corral *et al.* [14] with a category hypothesis. Forsyth [15] notice that there is a corresponding limited set of color signals for a given illumination and propose Gamut Mapping algorithm (GM). By computing the color signals under a certain illumination, the corresponding limited set, referred to as the canonical gamut, can be learned and used to constrain the possible illumination set for an input image. Gijsenji *et al.* [16] extend GM to generalized GM by using image derivative structures. Another classic supervised method is Bayesian color constancy (BCC) [17], which models the reflectance and illumination as random variables and estimates illumination from the posterior distribution. BCC is further extended by Gehler *et al.* [18]. Neural networks (NN)-based method [19] estimates illumination using a multi-layer perception fed with binary chromaticity histograms of images and corresponding illumination chromaticities. Better results have been achieved by using support vector regression (SVR) [20] that minimizes the structure risk without knowing the expected distribution of the image data. Cheng *et al.* [21] exploit four simple features with regression trees. Spatio-spectral statistics method (SSS) [22] is another efficient maximum likelihood approach that develops a statistical model for the spatial distribution of colors. A corrected moment illumination estimation algorithm (CM) [23] learns a regression mapping matrix of the color moments and the illumination. Focusing on natural image statistics (NIS), Gijsenij and Gevers [24] notice that image scenes can be classified using a Weibull distribution. Based on the scene classification, the optimal unsupervised method can be chosen for images in each scene category. Similarly, Image-classification-guided combination (IC) [25] provides a framework to select the most suitable unsupervised method using a decision forest for each image. And 3D stage geometry (SG) [26] models are used to determine the best color constancy method for different geometrical regions found in images. Exemplar-based color constancy [27] focuses on surfaces in the images. Nearest neighbor surfaces for each surface in a test image are found and illumination is estimated based on comparing the statistics of pixels belonging to nearest neighbor surfaces and the target surface. Bianco *et al.* [28] find that there are significant differences in content and illumination conditions among indoor and outdoor images. So, a prior indoor/outdoor (IO) classification is implemented to improve the performance of illumination estimation. Weijer *et al.* [29] improve illumination estimation using high-level-information (HVI). Images are modeled as a mixture of semantic classes. Then illumination estimation is guided by prior knowledge of the world, such as green grass, blue sky and gray roads. In [30], color statistics extracted from the faces are exploiting and clusters formed by skin colors are used as cues to estimate the illumination.

Recently, with the rapid growth of deep convolutional neural networks, several approaches have been proposed to solve the illumination estimation problem in an end-to-end manner and promising performance is demonstrated. One of the advantages of CNN is that it can take color images as input and incorporate feature learning into the training process. With a deep structure, CNN can learn complicated mappings while requiring minimal domain knowledge. Several typical methods include the works from Bianco *et al.* [31], [32], Hu *et al.* [33], Buzzelli *et al.* [34], Oh and Kim [35], Shi *et al.* [36], Barron [37] and Barron and Tsai [38].

Supervised methods are shown to be more accurate than unsupervised ones [39]. Because of the training phase with various categories of real-world scenes, supervised methods

are more robust and adaptable for different situations. However, supervised methods, especially the deep convolutional neural networks-based methods, should face a serious issue that is the lack of large scale of color image datasets of real-world scenes with ground-truth light color. Some methods try to solve this critical problem by implementing a training phase on synthetic datasets or public datasets without ground-truth illumination. Although these alternative datasets can be used, they do not reflect the complexity of realistic photometric effects and illumination in natural scenes. As we know, it is difficult and expensive to collect a large number of natural images of various scenes with different lighting conditions and measure corresponding illumination values.

III. Semi-Supervised Color Constancy

This section proposes a graph-based semi-supervised color constancy (SCC) method by using both labeled data (i.e., images with the ground-truth lighting colors) and unlabeled data (i.e., images without the ground-truth). Firstly, we give out a brief overview of the SCC. Then we discuss the feature extraction and graph construction for SCC.

A. Formulation

Assume that we are given n samples, which contains l training images $I_1, \ldots, I_l$ along with their corresponding ground-truth illumination chromaticities $\mathbf{c}_1, \ldots, \mathbf{c}_l$ ($\mathbf{c}_i = [c_r, c_g]^T, c_r = \log_2(G/R), c_g = \log_2(G/B)$) of the measured scene illumination and u training images $I_{l+1}, \ldots, I_n, (n = l + u)$ without the ground-truth. The visual feature of each training image I_i is represented as $\mathbf{x}_i \in \mathbb{R}^d, (i = 1, \ldots, n)$. The goal of graph-based SCC is to learn an illumination prediction function $\mathrm{f}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ using both labeled and unlabeled samples. To this end, inspired by the basic idea of semi-supervised learning, the function $\mathrm{f}(\mathbf{x})$ should have two major properties: (1) For those labeled samples, function $\mathrm{f}(\mathbf{x})$ should ensure that $\mathrm{f}(\mathbf{x}_i) \approx \mathbf{c}_i$, meaning that our estimated values are close to the given ground-truth; (2) All training samples, including unlabeled or labeled ones, should result in a similar estimation if they share certain similar visual characters. This motivates us to minimize the following loss function as:

$$\min \frac{1}{2}\left[\sum_{i=1}^{l}\|\mathrm{f}(\mathbf{x}_i) - \mathbf{c}_i\|^2 + \lambda_1 \sum_{i,j=1}^{n} s_{i,j}\left\|\mathrm{f}(\mathbf{x}_i) - \mathrm{f}(\mathbf{x}_j)\right\|^2 + \lambda_2\|\mathbf{w}\|^2\right], \tag{3}$$

where λ_1 is a hyperparameter which balances the contribution between the illumination estimation and image similarity. A graph represented by an adjacency matrix, $\mathbf{S} = [s_{i,j}], (i, j = 1, \ldots, n)$, indicates the similarity between two images I_i and I_j. The graph ensures that samples with similar features result in similar estimation results. A regularization term $\|\mathbf{w}\|^2$ with a coefficient λ_2 is added to control the model complexity and prevent overfitting. There are two important components in (3): (1) visual feature extraction, $\mathbf{x}_i \in \mathbb{R}^d$, for illumination estimation; (2) adjacency matrix $\mathbf{S}$ in graph construction. The following sections will detail these two parts.

B. Feature Extraction for SCC

The extraction of the feature vector $\mathbf{x}_i \in \mathbb{R}^d$ in (3) is used to map the relationship between an image and lighting color incident on it. Binarized chromaticity histogram, color moments and other visual features are widely used for illumination estimation. In this paper, inspired by recent work [38], the histogram based on a logarithm of the RGB color signal is used. We define a logarithm chromaticity space $(r, g)^T$ on RGB color space as:

$$\begin{cases} r = \log_2(G/R) \\ g = \log_2(G/B). \end{cases} \tag{4}$$

The range of R/G/B color signal depends on the format of the images, so the width of the histogram varies. Taking the SFU dataset [40] for instance that will be used in the following experiment, all images are in 8-bits data and the histogram is built on the data between $[-8, 8]$. We equally divide the r or g component into N bins and create a normalized histogram as the feature vector $\mathbf{x}_i \in \mathbb{R}^{N^2}$.

C. Graph Construction for SCC From Different Cues

The second key component in (3) is the adjacency matrix $\mathbf{S}$, indicating that images with similar visual characters should have similar illumination values. The loss function with the adjacency matrix makes our methods be an integration of regression and classification. Consequently, we create a graph $\mathbf{G}(\mathbf{V}, \mathbf{S})$ where the nodes $\mathbf{V}$ are all the images, represented by the feature vectors $[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ based on the selected cue. The edge between nodes (i, j) represents their similarity and the weight matrix $\mathbf{S}$ is computed as:

$$s_{i,j} = \exp\left(-\sum_{m=1}^{M}\frac{\left(v_{im} - v_{jm}\right)^2}{\sigma_m^2}\right), \tag{5}$$

where v_{im} is the m-th component of the feature vector $\mathbf{v}_i$ of the image I_i used in graph construction, σ_m is the corresponding scale hyperparameter for each component, and all components can share a same value.

How to select the feature vector $\mathbf{v}_i$ is the key step for graph construction. Inspired by the observations in [39], we can define feature vector $\mathbf{v}_i$ and construct corresponding graphs from low-, mid-, and high-level cues.

1) Low-Level Graph Using Color Distribution Cue: As we know, images having similar color distributions tend to be captured under similar illumination colors. Several different illumination estimation methods rely on this observation, including those based on neural networks [19], support vector regression [20], and color by correlation [13]. For low-level (LL) cues, images are treated as a bag of pixels, so LL cues are often the statistical distributions of color signals. Here we use the same histogram feature defined in section II-B, i.e. $\mathbf{v}_i = \mathbf{x}_i$.

2) Mid-Level Graph Using Initial Estimate Cue: Although the unsupervised color constancy methods are simple and inaccurate, the estimates of them can reflect the illuminates of images to a certain degree. So, we assume that images having similar initial estimates using simple unsupervised methods

tend to be captured under similar illumination colors. To this end, the grey edge framework in [8] is used to get initial estimates as:

$$\left(\int \left|\frac{\partial^t I^\sigma(\varepsilon)}{\partial \varepsilon^t}\right|^p dx\right)^{\frac{1}{p}} = h\mathbf{e}^{t,p,\sigma}, \tag{6}$$

where $I^\sigma(\varepsilon) = I(\varepsilon) \otimes G^\sigma$ is a convolution of the image with a Gaussian filter G^σ, t is the order of the derivative, ℓ is the Minkowski-norm and σ is the scale parameter of a Gaussian filter. With 9 different parameters settings of (t, ℓ, σ) based on the previous studies [24], [41], which are $(0, 1, 3)$, $(1, 2, 1)$, $(2, 1, 2)$, $(0, 1, 0)$, $(0, \infty, 0)$, $(0, 6, 0)$, $(2, 1, 5)$, $(0, 13, 2)$, $(1, 1, 6)$, initial illumination estimations are obtained. After changing into logarithm form using (4), an 18-dimensional feature vector is generated as our mid-level cue, which can be represented as $\mathbf{v}_i = [\mathbf{c}_i^{0,1,3}, \mathbf{c}_i^{1,2,1}, \mathbf{c}_i^{2,1,2}, \mathbf{c}_i^{0,1,0}, \mathbf{c}_i^{0,\infty,0}, \mathbf{c}_i^{0,6,0}, \mathbf{c}_i^{2,1,5}, \mathbf{c}_i^{0,13,2}, \mathbf{c}_i^{1,1,6}]^T$.

3) High-Level Graph Using Scene Category Cue: For a specific scene category, some corresponding lighting conditions are more likely to occur and this relationship can be exploited in illumination estimation. For example, compared with the skylight, indoor illumination is generally redder and more various. HL cues focus on those features that are reflections of the scene content and attempt to use this knowledge to guide illumination estimation. The Weibull parameterization is implemented to get our HL cue because it is valuable in determining the scenes of an image. Further, we can fit the distribution of edge responses with a Weibull distribution:

$$wb(a) = \vartheta \exp\left(-\frac{1}{\alpha}\left|\frac{z}{\beta}\right|^\alpha\right), \tag{7}$$

where α and β reflect the grain size and image contrast respectively. z represents the edge responses and ϑ is a constant for normalization. With the same parameter setting in [24], we transform the image into the 3-dimensional opponent color space (O_1, O_2, O_3)[1] and a 24-dimensional feature vector can be obtained as our HL cue.

IV. Multi-Cue Semi-Supervised Color Constancy

In this section, we extend SCC to integrate three cues and optimize the corresponding objective function. According to Section II, we can obtain three different adjacency matrices, $\mathbf{S}^k = [s_{i,j}^k]$, $(i, j = 1, \ldots, n; k = 1, 2, 3)$, with different cues, resulting in different SCC models: SCC with low-level graph (SCC_L), SCC with mid-level graph (SCC_M), and SCC with high-level graph (SCC_H). However, any cue can only reflect some constraints from one viewpoint. In order to obtain more accurate and stable illumination estimation, we integrate these three cues into a unified semi-supervised learning model by extending SCC to multi-cue semi-supervised color constancy (MSCC).

A. Formulation of MSCC

The objective function of MSCC includes three aspects:

$$\mathrm{J} = \mathrm{J}_{label} + \lambda_1 \mathrm{J}_{unlabel} + \lambda_2 \mathrm{J}_{uni}. \tag{8}$$

[1] $O_1 = \frac{R-G}{\sqrt{2}}$, $O_2 = \frac{R+G-2B}{\sqrt{6}}$, $O_3 = \frac{R+B+G}{\sqrt{3}}$

Similar to SCC, the first term J_{label} defines the similarity between the estimation and ground-truth for those labeled data on three cues:

$$\mathrm{J}_{label} = \sum_k \sum_{i=1}^{l} \|\mathrm{f}_k(\mathbf{x}_i) - c_i\|^2. \tag{9}$$

Here, for each cue, f_k is defined as a linear regression function $\mathrm{f}_k(x) = \mathbf{w}_k^T \mathbf{x}$ $(k = 1, 2, 3)$, which represents the illumination prediction function of SCC_L, SCC_M, SCC_H, respectively. $\mathbf{w}_k^T = [w_k^1, \ldots, w_k^d] \in \mathbb{R}^d$ is the corresponding weight vector. All these 3 weights' vectors can be concatenated into a matrix, $\mathbf{W} \in \mathbb{R}^d \times k$, to be estimated.

The second term focuses on the distance consistency among samples, which is the key idea of the graph-based semi-supervised method. So we have $\mathrm{J}_{unlabel}$:

$$\mathrm{J}_{unlabel} = \sum_k \left(\sum_{i,j=1}^{n} S_{i,j} \left\|\mathrm{f}_k(\mathbf{x}_i) - \mathrm{f}_k(\mathbf{x}_j)\right\|^2 + \delta \|\mathbf{w}_k\|^2\right). \tag{10}$$

The last term J_{uni} defines a consistency constraint on different distance matrices from these three cues, that is, the different $\mathbf{w}_k$ from different cues should share similar values. So we have:

$$\begin{aligned} \mathrm{J}_{uni} &= \sum_{k=1}^{3} \|\mathbf{w}_k - \bar{\mathbf{w}}\|^2, \\ \bar{\mathbf{w}} &= \sum_{k=1}^{3} \mathbf{w}_k / 3. \end{aligned} \tag{11}$$

where $\bar{\mathbf{w}}$ is the mean vector of these three models.

B. Optimization

Borrowing the basic ideas from [42], we develop an efficient and simple method to solve the loss function in (8). In the first term J_{label}, for each cue k, the corresponding $\mathbf{w}_k$ can be further represented as a linear combination of the coefficient vector $\mathbf{h}_k = [h_k^1, h_k^2, \ldots, h_k^n]^T$ and feature data, that is $\mathbf{w}_k = \sum_{i=1}^{n} h_k^i \mathbf{x}_i = \mathbf{X}_k \mathbf{h}_k$. Then a label truncated identity matrix $\mathbf{J}_k \in \mathbb{R}^{n \times n}$ and an initial label vector $\mathbf{b}_k \in \mathbb{R}^{n \times 1}$ for each cue's weight vector, along with some notations are defined as follows:

$$\begin{aligned} J_{ii} &= \begin{cases} 1, & i \in l \\ 0, & otherwise, \end{cases} \\ b_k(i) &= \begin{cases} 1, & i \in l \\ 0, & otherwise, \end{cases} \\ \mathbf{X} &= [\mathbf{X}_1, \ldots, \mathbf{X}_k], \\ \mathbf{H} &= diag(\mathbf{h}_1, \ldots, \mathbf{h}_k), \\ \mathbf{B} &= diag(\mathbf{b}_1, \ldots, \mathbf{b}_k), \\ \mathbf{A} &= diag(\mathbf{A}_1, \ldots, \mathbf{A}_k), \\ \mathbf{J} &= diag(\mathbf{J}_1, \ldots, \mathbf{J}_k), \end{aligned} \tag{12}$$

where $\mathbf{A}$ is the Gram matrix of $\mathbf{X}$. *diag* means that we arrange vectors or matrices along the diagonal line and form

a new matrix. Now J_{label} can be rewritten as $\mathrm{J}_{label} = \|\mathbf{B} - \mathbf{JAH}\|_F^2$. Also, for the second term $\mathrm{J}_{unlabel}$ in the loss function, we have $\mathrm{J}_{unlabel} = tr\left(\mathbf{H}^T\left(\mathbf{A} + \delta\mathbf{ALA}\right)\mathbf{H}\right)$. For the last term J_{uni}, we introduce a vector $\mathbf{E} = (1,1,1)^T$ and $\mathbf{M} = \mathbf{I} - \mathbf{E}\left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T$. Then we have $\mathrm{J}_{uni} = tr\left(\mathbf{WMW}^T\right) = tr\left(\mathbf{XHMH}^T\mathbf{X}^T\right)$. Now we can rewrite our loss function J as:

$$\mathrm{J} = \|\mathbf{B} - \mathbf{JAH}\|_F^2 + \lambda_1 tr\left(\mathbf{H}^T\left(\mathbf{A} + \delta\mathbf{ALA}\right)\mathbf{H}\right) + \lambda_2 tr\left(\mathbf{XHMH}^T\mathbf{X}^T\right). \quad (13)$$

By optimizing $\mathbf{H}$ and setting $\frac{\partial J}{\partial \mathbf{H}} = 0$, we can get:

$$\left(\lambda_1\mathbf{A} + \mathbf{A}\left(\mathbf{J} + \lambda_1\delta\mathbf{L}\right)\mathbf{A}\right)\mathbf{H} + \lambda_2\mathbf{X}^T\mathbf{XHM} = \mathbf{AB}. \quad (14)$$

The equation can be easily solved by using the method in [43]. We can also apply gradient descent to solve $\mathbf{H}$ instead if the scales are too large to solve directly.

C. Combination of Estimates From Multiple Cues

We can obtain three estimates using MSCC with three cues for an image. We denote the estimate with low-level graph as $f_L(\mathbf{x})$, the estimate with low-level graph as $f_M(\mathbf{x})$, and the estimate with high-level graph as $f_H(\mathbf{x})$. The final estimate can be computed using a weighted average, as:

$$\hat{\mathbf{c}}_i = \alpha' f_L(\mathbf{x}_i) + \beta' f_M(\mathbf{x}_i) + (1 - \alpha' - \beta') f_H(\mathbf{x}_i), \quad (15)$$

where $\alpha', \beta' \in [0,1]$ and the optimal values are determined by a simple exhaustive search method on the training set in this paper.

V. Experiment

We evaluate the performance of our proposed methods and other well-known methods on three datasets: Gehler-Shi [18], NUS [10] and Linear SFU image set [40]. The first one is provided by Gehler *et al.* [18] and then is reprocessed by Shi and Funt [44]. The second one is from Cheng and Brown [10], which is also a common dataset of real-world images and illuminations. The third one is produced by Ciurea and Funt [40] from a digital video and is linearized by Gijsenij *et al.* [16]. In the experiments, we implement our methods using following parameters. In the feature extraction process, the number of bins, N, is set as 4096, which generates a $64 * 64$ histogram. In the training process, λ_1, λ_2 in (8), δ in (10) are set as 10^2, 10^{-1} and 10^{-2} respectively. In the combination of estimates from multiple cues, α' and β' in (16) are determined by a grid search. α' is set as 0.5 and β' is set as 0.4. The σ in (5) varies with cues we use. σ for LL and ML cue is set as 0.01, and for HL cue, σ is set as 0.001.

A. Error Measurement

Angular error is chosen to evaluate the methods. Given an estimated chromaticity $\mathbf{c} = \left[c_r, c_g\right]^T$, we can calculate color signal $\mathbf{e} = \left[e_r, e_g, e_b\right]$ as:

$$e_r = \frac{2^{(-c_r)}}{\varsigma}, \quad e_g = \frac{1}{\varsigma}, \quad e_b = \frac{2^{(-c_g)}}{\varsigma}, \quad \varsigma = \sqrt{2^{(-c_r)^2} + 2^{(-c_g)^2} + 1}. \quad (16)$$

For each image, the ground-truth light source $\mathbf{e}_a$ is known and the estimated illumination $\mathbf{e}_y$ can be gained from each color constancy method. To evaluate how close $\mathbf{e}_y$ resembles $\mathbf{e}_a$, the angular error can be computed as follows:

$$angular(\mathbf{e}_y, \mathbf{e}_a) = \frac{180^\circ}{\pi}\cos^{-1}\left(\frac{\mathbf{e}_y \bullet \mathbf{e}_a}{\|\mathbf{e}_y\|\,\|\mathbf{e}_a\|}\right), \quad (17)$$

where $\bullet$ is the dot product and $\|\|$ indicates the Euclidean norm. Besides the mean and the median of the angle errors, the tri-mean, the mean of the best 25% (B-25%) and the mean of the worst of 25% (W-25%) angle errors are also provided for a more overall comparison. The best 25% (or worst 25%) measures the mean of the smallest (or largest) 25% angle errors on the test images. The tri-mean is calculated as the average of the median and the midhinge.

B. Experiments Without Unlabeled Training Samples

In this section, experiments are conducted without unlabeled training samples on three datasets. The performance of our proposed methods and other well-known methods is reported and analyzed. Without unlabeled samples, our proposed method works as a multi-task supervised method.

1) Results on the Gehler-Shi Set: The Gehler-Shi dataset contains 568 images taken using two high-quality DSLR cameras (Canon 5D and 1D). The dataset includes a wide variety of both indoor and outdoor scenes. Each image in the dataset contains a color checker, which performs as a calibration object for the calculation of the ground-truth illumination. The original images in the dataset are saved as Canon RAW format and this brings the problem of clipped pixels, which are non-linear and include the effect of the camera's white balancing. To solve the problem, reprocessing is taken by Shi *et al.* [18] and a new dataset containing almost raw 12-bit PNG format images is produced. The images are named in the sequence in which they were taken. As a result, neighboring images in the sequence are more likely than others to be taken from similar scenes. To ensure that the scenes from the training set and the test set have no overlap, we use an uncorrelated threefold cross validation provided by Li *et al.* [41]. In this experiment, both the SCC and the MSCC use the full training set without unlabeled samples, which is the same as the other supervised methods. Besides MSCC, methods combining two cues are exploited as comparisons. We represent the method containing low- and mid-level cues as SCC_LM, the method containing low- and high-level cues as SCC_LH and the method containing mid- and high-level cues as SCC_MH.

We compare the proposed methods with 29 previous methods, including both supervised and unsupervised methods. The results are collected from colorconstancy.com [45]. According to the experimental results in Table I, the proposed methods, both SCC and MSCC, outperform all the unsupervised methods and most supervised methods. MSCC outperforms all the methods listed in Table I except DS-net [36] and FFCC [38]. DS-net is a deep learning based method that needs a long time training procedure on the high-performance computers with expensive GPUs, while MSCC can be efficiently trained on common PCs. It should be noted that, the error values of the

TABLE I
PERFORMANCE COMPARISON OF OUR PROPOSED METHODS AGAINST OTHER METHODS ON THE GEHLER-SHI DATA SET

Method		Mean	Median	Trimean	B-25%	W-25%
Unsupervised	GW [6]	6.40	6.30	6.30	-	-
	SoG [7]	4.90	4.00	4.20	-	-
	MaxRGB [4]	7.60	5.70	6.40	-	-
	GE1,1,6 [8]	5.30	4.50	4.70	-	-
	Zhang et al. 2016 [9]	4.80	2.70	-	-	-
	Yang et al. 2015 [11]	4.60	3.10	-	-	-
	Cheng et al. 2014 [10]	3.52	2.14	2.47	0.50	8.74
	Bianco et al. 2019 [12]	3.46	2.23	-	-	-
Supervised	NN [19]	5.13	3.77	4.06	1.12	11.50
	SVR [20]	4.08	3.23	3.35	0.72	9.03
	SSS [22]	3.96	3.24	3.46	1.52	7.61
	GM [16]	5.96	3.98	4.53	0.83	14.30
	Bayesian [18]	4.82	3.46	3.88	1.26	10.49
	SVRC [20]	3.55	2.52	2.73	0.32	8.24
	NIS [24]	4.20	3.18	3.49	0.54	10.30
	IC [25]	3.87	2.83	3.07	0.36	9.29
	IO [28]	5.18	3.55	4.00	0.65	12.40
	SG [26]	4.49	3.09	3.45	0.56	10.60
	HVI [29]	4.31	3.06	3.38	0.80	9.86
	MC [39]	3.25	2.20	2.55	0.30	8.13
	Cheng et al. 2015 [21]	2.45	1.65	1.75	0.38	5.87
	CCC [37]	1.95	1.22	1.38	0.35	4.76
	FFCC [38]	1.61	0.86	1.02	0.23	4.27
	FFCC#	2.91	1.98	2.25	0.65	6.86
	Exemplar-based [27]	2.89	2.27	2.42	0.82	5.07
	Bianco et al. 2015 [31]	2.63	1.89	-	-	-
	FC4 [33]	1.65	1.18	1.27	0.38	3.78
	Buzzelli et al. 2018 [34]	4.84	4.12	-	-	-
	Oh et al. 2017 [35]	2.16	1.47	-	-	-
	DS-net [36]	1.90	1.12	-	-	-
Semi-supervised	SCC_L	2.29	1.32	1.61	0.33	7.09
	SCC_M	2.39	1.38	1.75	0.34	6.73
	SCC_H	3.16	1.78	2.63	0.62	8.31
	SCC_LM	2.08	1.21	1.52	0.31	6.51
	SCC_LH	2.19	1.30	1.57	0.32	7.06
	SCC_MH	2.39	1.38	1.71	0.36	6.40
	MSCC	2.05	1.15	1.45	0.29	6.02

FFCC are based on random threefold cross validation on this dataset, not our uncorrelated threefold cross validation. The results of FFCC trained on related threefold cross validation are shown in Table I as FFCC#, which is trained using the code provided in [38] and has much lower performance than FFCC. Besides, we can find that the MSCC outperforms all the SCC methods. The reason for the improvement is mainly due to two reasons: (1) the combination step of different cues makes the results stable and accurate, and more importantly (2) the key novelty, that all training samples should result in a similar estimation if they share certain similar visual characters, ensures that different cues can lead to a better estimated illumination. The SCC_H achieves lower performance than both SCC_L and SCC_M, which implies that the scene category cue is a loose constraint for illumination estimation. All the results indicate that the proposed semi-supervised learning-based methods are better alternatives than some supervised learning methods for illumination estimation, even with no unlabeled samples. Fig.2 shows several visual comparison samples of corrected images using different methods.

Because of the problem raised by G.D.Finlayson [46]–[48], experiments based on the REC ground-truth are also implemented. The REC ground-truth notices the problem in calculating correctly the bounding boxes for Shi's release. Results of 12 previous methods are collected from colorconstancy.com and shown in Table II. It can be seen that our methods still achieve comparable results under the REC ground-truth. SCC methods can obtain leading performance, and the combination of different cues can further reduce the angular error in most cases. It should be noticed that the performance of SCC_H is less satisfactory, as the semantic information may be unreliable with a wide variety of scenes, which also results in unsatisfactory performance of SCC_MH.

2) Results on the NUS Set: The NUS set is produced by Cheng and Brown [10] and is composed of 1736 high-quality images. The images are taken from 8 commercial cameras (Canon 1DS Mark III, Canon 600D, Fujifilm XM1, Nikon D5200, Olympus EPL6, Panasonic GX1, Samsung NX 2000, and Sony α57). For each camera, over 200 images, which contains both indoor and outdoor scenes, are captured. Similar to the Gehler-Shi set, a color checker in each image is used to provide the ground-truth illumination.

We report the results from [37] for 22 methods. For each camera, learning-based methods are trained and tested separately, and a threefold cross validation is taken. Then the average results of 8 cameras are reported in Table III as the evaluation criterion. It shows that our proposed semi-supervised methods achieve comparable performance

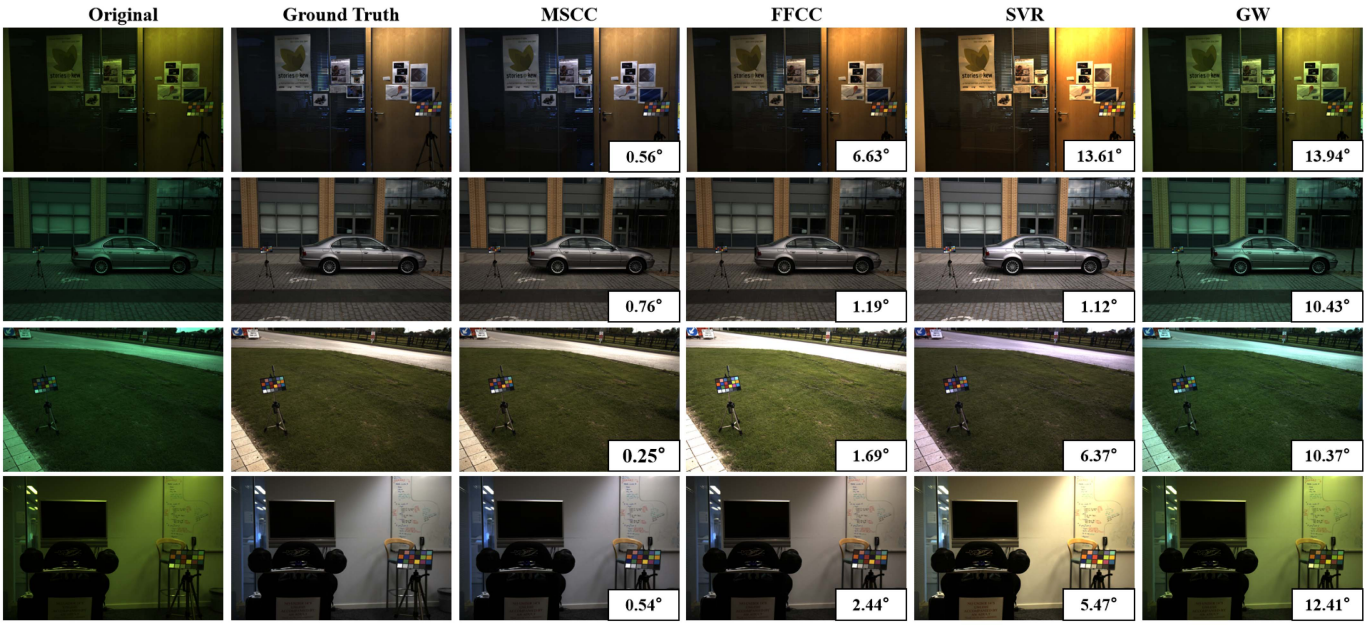


Fig. 2. Some examples of corrected images using different methods. The angular error is shown in the lower right corner of each image.

TABLE II

PERFORMANCE COMPARISON OF OUR PROPOSED METHOD AGAINST VARIOUS OTHER METHODS ON THE G.HEMRIT'S COLOR CHECKER DATASET

Method		Mean	Median	Trimean	Max
Unsupervised	GW [6]	9.7	10.0	10.0	24.8
	MaxRGB [4]	9.1	6.7	7.8	43.0
	SoG [7]	7.3	6.8	6.9	22.5
	2nd-order-GE [8]	4.0	3.1	3.3	20.6
	General GW [2]	6.6	5.9	6.1	23.0
Supervised	SVR [20]	11.0	9.6	10.1	32.5
	Bayesian [18]	5.4	3.8	4.3	25.5
	NIS [24]	5.6	4.7	4.9	30.6
	IC [25]	6.1	5.1	5.3	24.7
	Exemplar-based [27]	4.9	4.4	4.6	14.5
	Bianco et al. 2015 [31]	7.0	5.3	5.7	29.1
	FFCC [38]	2.0	1.1	1.4	25.0
Semi-supervised	SCC_L	2.4	1.4	1.5	20.9
	SCC_M	2.5	1.4	1.7	18.9
	SCC_H	3.1	1.8	2.4	23.0
	SCC_LM	2.0	1.2	1.5	14.9
	SCC_LH	2.2	1.3	1.7	15.3
	SCC_MH	2.9	1.5	1.4	16.5
	MSCC	2.0	1.2	1.3	14.6

with the state-of-art color constancy methods. Besides, MSCC outperforms most supervised methods and is more stable because of the combination of different cues, which indicates that more reliable results can be gained in various situations with MSCC. In addition, the performance of MSCC is much better than deep learning-based method DS-Net. The reason is that the NUS set only contains around 200 images for each camera. It is difficult to learn a good deep neural network model using such a small scale set.

3) Results on the Linear SFU Set: The SFU dataset is provided by Ciurea and Funt [40] and contains 11346 images. The images are from the frames of videos, so a high relevance exists. To solve this problem, a fixed version of the dataset is applied by Bianco *et al.* [28]. 1135 less correlated images are selected according to a video-based analysis.

To further avoid the correlation, we divide the dataset into 15 parts according to the files they belong to, which indicate the scenes where the images are taken. Then we apply a 15-fold cross validation. Each file is regarded as the test set separately and the other 14 files are regarded as the training set. This strategy ensures that the training set and test set are truly distinct. We report the results from [39] for 15 methods. According to Table IV, all the proposed SCC and MSCC methods outperform all the other methods including both unsupervised and supervised ones. Especially, MSCC reduces the median and mean errors by 24% and 18% respectively comparing with the best supervised method MC.

C. Experiment With Unlabeled Training Samples

This section shows the experiments on different datasets by gradually decreasing the number of labeled images.

TABLE III
PERFORMANCE COMPARISON OF OUR PROPOSED METHODS AGAINST OTHER METHODS ON THE NUS DATA SET

Method		Mean	Median	Trimean	B-25%	W-25%
Unsupervised	MaxRGB [4]	10.62	10.58	10.19	1.86	19.45
	GW [6]	4.14	3.20	3.39	0.90	9.00
	SoG [7]	3.40	2.57	2.73	0.77	7.41
	General GW [2]	3.21	2.38	2.53	0.71	7.10
	2nd-order-GE [8]	3.2	2.26	2.44	0.75	7.27
	1st-order-GE [8]	3.2	2.22	2.43	0.72	7.36
	Bianco et al. 2019 [12]	3.00	2.25	-	-	-
Supervised	CM [23]	3.05	1.90	2.13	0.65	7.41
	Edge-based GM [16]	8.43	7.05	7.37	2.41	16.08
	Pixel-based GM [16]	7.70	6.71	6.90	2.51	14.05
	IC [25]	7.20	5.96	6.28	2.20	13.61
	Bayesian [18]	3.67	2.73	2.91	0.82	8.21
	Buzzelli et al. 2018 [34]	4.32	3.37	-	-	-
	NIS [24]	3.71	2.60	2.84	0.79	8.47
	SSS(ML) [22]	3.11	2.49	2.60	0.82	6.59
	SSS(GenPrior) [22]	2.96	2.33	2.47	0.80	6.18
	Cheng [21]	2.92	2.04	2.24	0.62	6.61
	CCC [37]	2.38	1.48	1.69	0.45	5.85
	FFCC [38]	1.99	1.31	1.43	0.35	4.75
	FC4 [33]	2.23	1.57	1.72	0.47	5.15
	Oh et al. 2017 [35]	2.41	2.15	-	-	-
	DS-net [36]	2.24	1.46	-	-	-
Semi-supervised	SCC_L	2.41	1.55	2.07	0.53	6.97
	SCC_M	2.26	1.58	2.06	0.61	6.24
	SCC_H	2.90	1.91	2.69	0.64	7.80
	SCC_LM	2.15	1.38	1.87	0.53	6.06
	SCC_LH	2.33	1.43	1.95	0.55	6.63
	SCC_MH	2.21	1.44	1.88	0.55	5.73
	MSCC	2.15	1.35	1.84	0.53	5.51

TABLE IV
PERFORMANCE COMPARISON OF OUR PROPOSED METHODS AGAINST OTHER METHODS ON THE SFU DATA SET

Method		Mean	Median	Trimean	B-25%	W-25%
Unsupervised	GW [6]	13.00	10.80	11.30	3.24	26.20
	SoG [7]	11.60	10.40	10.60	3.52	22.10
	MaxRGB [4]	12.70	10.30	11.30	2.26	26.40
	GE1,1,6 [8]	11.10	9.15	9.70	3.06	22.10
Supervised	NN [19]	11.80	9.75	10.20	3.21	23.80
	SVR [20]	9.98	8.39	8.74	2.74	20.00
	SSS [22]	10.40	8.74	9.20	3.04	20.60
	GM [16]	14.20	12.00	12.70	3.13	28.60
	SVRC [20]	9.39	7.54	8.02	2.25	19.50
	NIS [24]	10.90	8.16	9.15	2.24	23.70
	IC [25]	10.20	7.71	8.57	2.18	22.20
	IO [28]	10.30	7.73	8.70	2.07	22.50
	SG [26]	10.80	8.93	9.52	2.49	21.90
	HVI [29]	10.80	8.51	9.06	2.59	23.00
	MC [39]	8.81	5.61	6.78	1.50	19.10
Semi-supervised	SCC_L	7.55	4.51	5.37	0.85	19.25
	SCC_M	8.09	4.75	6.01	0.96	20.71
	SCC_H	9.08	5.25	6.59	0.84	23.21
	SCC_LM	7.24	4.45	5.25	0.98	18.25
	SCC_LH	7.40	4.38	5.20	0.85	18.83
	SCC_MH	8.09	4.75	6.01	0.96	20.71
	MSCC	7.22	4.27	5.10	0.92	17.58

We further test the proposed semi-supervised methods with limited labeled training samples.

1) Results on the Gehler-Shi Set: The uncorrelated three-fold cross validation setting is also used in this experiment. Differently, in each cross validation, the training dataset is composed of two parts: $p\%$ data are randomly selected out as labeled images, while the remaining are used as unlabeled. Those selected labeled images are used to train SVR, NN and FFCC Methods. The whole training set, including labeled and unlabeled ones, are used to train our SCC and MSCC methods. The procedure is repeated 5 times and the mean performance is used as the final result for each method. Fig.4 shows the mean and median errors with different values of the percentage.

According to Fig.4, along with the reducing percentage, the error of SVR, NN and FFCC increases rapidly. Though the FFCC method achieves the best performance on the Gehler-Shi Set with $p = 100$, its performance varies dramatically with the change of p. Since the deep neural network cannot be

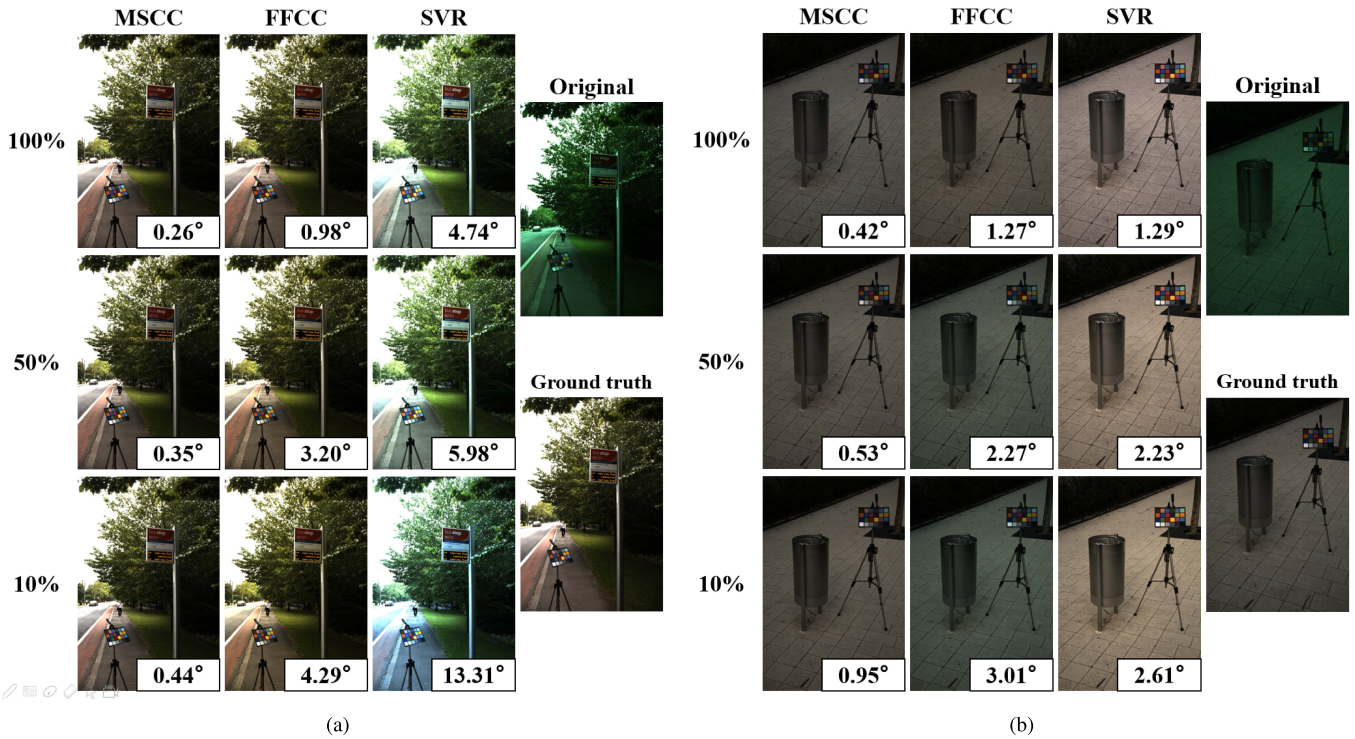


Fig. 3. Some corrected examples of different methods with different amounts of labeled training data. Our semi-supervised method achieves a stable performance when the number of labeled samples decreases.

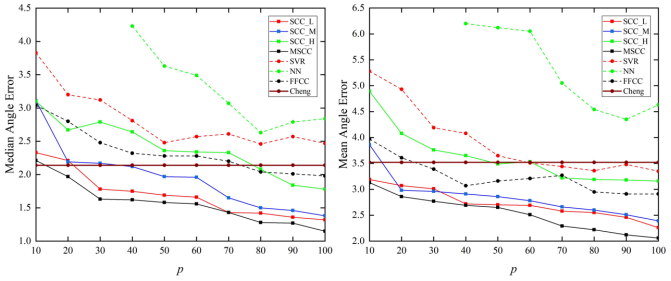


Fig. 4. Experiment results with few training samples on the Gehler-Shi Set. Cheng *et al.* 2014 [10] is shown as a baseline of unsupervised method. It shows that even with a small number of labeled samples, our proposed method can still perform well.

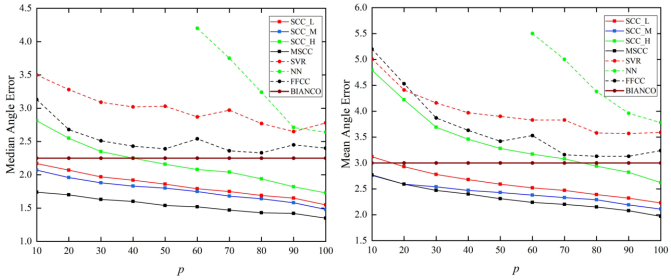


Fig. 5. Experiment results with few training samples on the NUS Set. Biano *et al.* 2019 [12] is shown as a baseline of unsupervised method. Performance of supervised methods drops sharply while facing insufficient labeled samples.

well trained on a small training set, we use a shallow neural network architecture in [19] instead. Even though the NN method has much fewer parameters than deep neural network, it also cannot produce a reasonable model when $p < 30$. In comparison, the MSCC still achieves high performance even when $p = 10$ (about 38 labeled images). In this situation, the median and mean errors of MSCC are 2.21° and 3.13°, which are much lower than most unsupervised methods and supervised methods with 100% labeled training samples in Table I. The SCC methods also obtain good performance with so few training samples. This indicates that low-level features become more and more important as the number of training samples decreases. Some corrected examples are shown in Fig.3.

2) Results on the NUS Set: Similar to the setting of the experiments on the Gehler-Shi set, we compare different methods with limited training samples on the NUS set. We implement several methods with their best parameter setting and Fig.5 shows the change of performance with different amounts of labeled training samples. Although the numbers of images for 8 camera sets are slightly different, we show the average of results of 8 cameras here for convenience. It should be noted that the number of training samples is as low as 20 in some camera datasets when $p = 10$. And our semi-supervised methods can still get reasonable results. Both MSCC and SCC_L perform better than all the unsupervised methods in all sets of the training set. The median error of MSCC increases by only 0.26° (from 1.35° to 1.61°) even when the number of labeled samples reduces by 90%. It further validates the stableness of the MSCC and SCC methods with limited training samples.

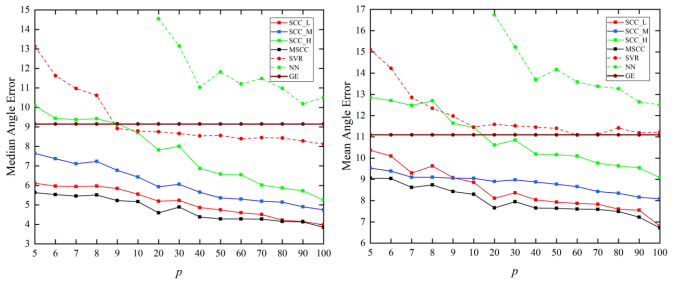


Fig. 6. Experiment results with few training samples on the SFU Set. GE [8] is shown as a baseline of unsupervised method. Although all the methods do not perform well because of the unusual lighting conditions in the dataset, our methods still achieve a leading performance.

3) Results on the Linear SFU Set: The linear SFU set contains more images than the Gehler-Shi set and NUS set, so we set the percentage range p between $[5, 100]$ to simulate the situation with limited training samples. The result is shown in Fig.6. It shows that our semi-supervised methods perform well while other learning methods do not. It is noted that NN becomes instable. The performance of SVR seems rather steady when $p > 9$. However when the number of labeled samples continues decreasing ($p < 9$), the performance of it declines sharply. On the contrary, the semi-supervised methods, especially MSCC, show high and stable performance, even when $p < 9$. The performance of SCC_H is less satisfactory. This indicates that only using the high-level semantic information may be unreliable when the training set is small but complex.

D. Comparison Among Different Cues

In this section, different methods generated by the MSCC framework with different cue combinations are compared and discussed. We compare results on the Gelher Shi set, with different amounts of labeled training samples, as shown in Fig.7. It is noticed that SCC_L achieves the best performance among methods based on the single cue. And SCC_M can perform as well as SCC_L in most cases. This indicates that low-level and mid-level features are meaningful even when there are only a few training samples. Meanwhile, SCC_H performs worse as the percent of training samples goes down. And the gap between SCC_H and other two single cue methods is getting larger. This is because that the scene understanding itself is a challenging problem, which calls for discriminative features and a number of training samples. Although the performance of SCC_H may be less satisfactory, it is noticed that MSCC methods with high level cue generally achieve better results. This means that high-level cues can be a useful supplement to low-level and mid-level cues.

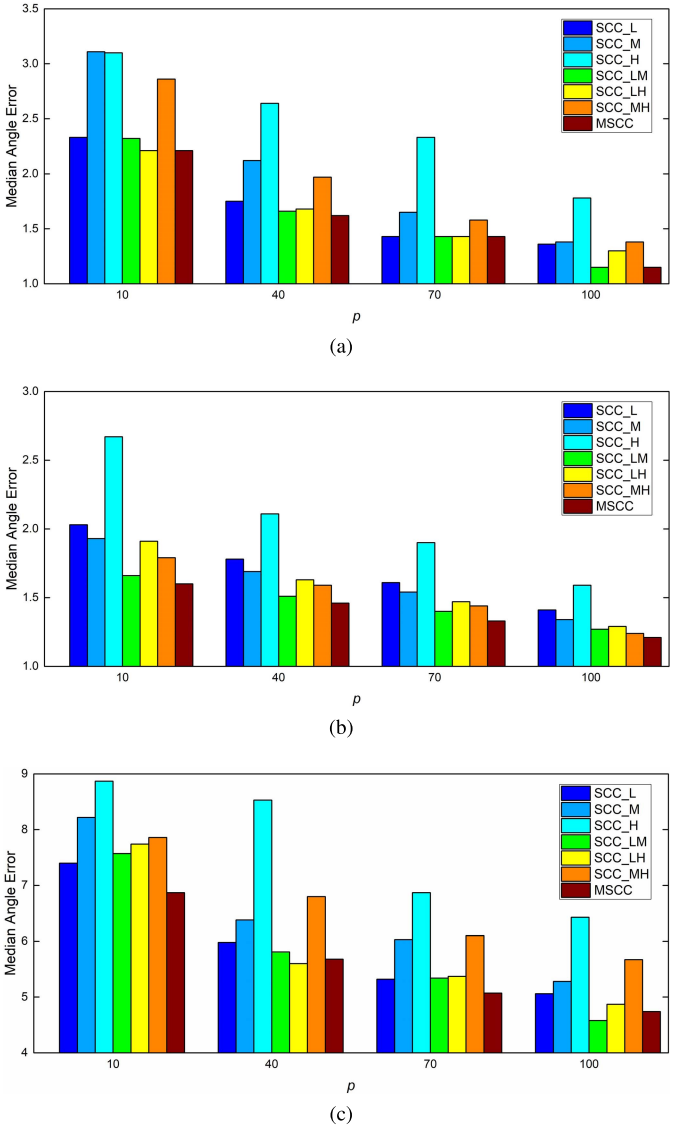


Fig. 7. Performance comparison of SCC_L, SCC_M, SCC_H, SCC_LM, SCC_LH, SCC_MH and MSCC. In each dataset, experiments with different amounts of labeled training samples are conducted. (a). Gehler set, (b). NUS set, (c). SFU set.

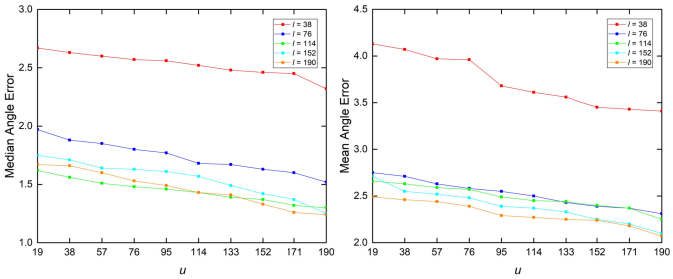


Fig. 8. Experiments with an increasing number of unlabeled samples. Situations with different amounts of labeled samples are reported.

E. The Effect of Unlabeled Set Size

In this section, we further discuss the effect of the number of unlabeled samples. First, we equally divide the dataset into three parts as labeled set, unlabeled set and test set respectively. Second, we randomly choose l samples from labeled set and u unlabeled samples from the unlabeled set to compose a training set. Finally, the SCC and MSCC are trained on the training set with l labeled samples and u unlabeled samples, and evaluated on the test set. The performance of SCC and MSCC with different values of l and u on the Gehler Shi set are reported in Fig.8. It can be seen that as the number

TABLE V
MEDIAN ERROR ANGLES OF EXPERIMENTS WITH AUGMENTED IMAGES

P	SCC_L	SCC_L#	LP(%)	SCC_M	SCC_M#	LP(%)	SCC_H	SCC_H#	LP(%)	MSCC	MSCC#	LP(%)
10	2.33	2.15	8	2.80	2.08	26	4.35	2.35	46	2.21	1.78	19
20	2.21	1.98	10	2.19	1.98	10	4.31	2.27	47	1.97	1.65	16
30	1.78	1.73	3	1.97	1.87	5	4.22	2.16	49	1.63	1.61	1
40	1.75	1.65	6	1.95	1.77	9	2.91	2.11	27	1.62	1.54	5
50	1.69	1.55	8	1.92	1.65	14	2.74	2.04	26	1.58	1.53	3
60	1.66	1.48	11	1.92	1.62	16	2.52	1.95	23	1.56	1.41	10
70	1.43	1.36	5	1.74	1.60	8	2.39	1.73	28	1.43	1.31	8
80	1.42	1.31	8	1.64	1.47	10	2.13	1.72	19	1.28	1.21	5
90	1.36	1.25	8	1.52	1.37	10	1.99	1.66	17	1.27	1.16	9
100	1.24	1.18	5	1.32	1.13	14	1.52	1.51	1	1.15	1.11	3

TABLE VI
MEAN ERROR ANGLES OF EXPERIMENTS WITH AUGMENTED IMAGES

P	SCC_L	SCC_L#	LP(%)	SCC_M	SCC_M#	LP(%)	SCC_H	SCC_H#	LP(%)	MSCC	MSCC#	LP(%)
10	3.19	2.94	8	4.41	3.35	24	6.80	5.17	24	3.13	2.82	10
20	3.07	2.77	10	3.16	3.00	5	6.09	3.46	43	2.86	2.63	8
30	3.01	2.73	9	3.06	2.98	3	6.05	3.19	47	2.77	2.39	14
40	2.72	2.61	4	2.95	2.81	5	5.67	3.18	44	2.69	2.28	15
50	2.70	2.53	6	2.83	2.75	3	5.48	3.06	44	2.65	2.19	17
60	2.69	2.51	7	2.80	2.68	4	5.45	2.88	47	2.51	2.18	13
70	2.58	2.48	4	2.79	2.66	5	5.28	2.83	46	2.29	2.15	6
80	2.55	2.47	3	2.70	2.59	4	5.12	2.81	45	2.22	2.13	4
90	2.46	2.38	3	2.59	2.48	4	4.58	2.80	39	2.12	2.06	3
100	2.26	2.00	12	2.50	2.44	2	4.07	2.60	36	2.06	1.87	9

of the unlabeled samples increases, both median and mean angular errors are decreasing. Though the advance caused by unlabeled samples are not as good as that caused by the same amount of labeled samples, it is still effective and practical. The unlabeled samples provide additional information for the distribution of the dataset. By exploiting the extended graph, which represents the relationship of both labeled samples and unlabeled ones, a more accurate illuminant estimation can be gained. This is the reason why the unlabeled samples can improve the illumination estimation accuracy. As we know, it is much easier to collect a large number of unlabeled samples, which makes the proposed semi-supervised methods effective and practical.

F. The Effect of Augmented Images

Inspired by [37], we extend the training data set after generating three extra images from each input image artificially. Let I represent the original image, the four augmented images can be derived from the following equations:

$$\begin{aligned} I_1 &= I, \\ I_2 &= \max\left(0, I * \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}\right), \\ I_3 &= blur\left(I^4, 11\right)^{1/4}, \\ I_4 &= \sqrt{blur\left(I^2, 3\right) - blur(I, 3)^2}, \end{aligned} \tag{18}$$

in which $blur$ is a box filter and the details can be found in [37] and $*$ represents a convolution operation. Using the same experimental settings in Section IV-B, we test the SCC and MSCC methods on different images (I_1, I_2, I_3, I_4) respectively with different p on the Gehler-Shi Set. Accordingly, we can obtain four estimates for each image I. Then the four estimates are averaged as the final estimation. The median and mean errors are shown in Tables V and VI. The methods without '#' mean using only original image I_1, while the methods with '#' mean combining estimates form all the four images. LP means the error reduction of the method using four augmented images comparing with the method using original image. From the comparison in Tables V and VI, the augmented images can provide much more information of an image from different views and this results in better performance, especially when the labeled training samples are limited. For example, when $p = 10$, the median and mean errors of MSCC# are reduced by 19% and 10% respectively comparing with MSCC. It indicates that combining different augmented images is an effective way to improve the performance of the semi-supervised methods when the labeled training samples are very limited.

G. Efficiency Comparison

In this section, the computational cost of different methods is summarized and compared. We use the average computational time for each image as the evaluation criterion. The Gehler-Shi dataset is chosen to implement the experiment. For unsupervised methods, no training process is needed and only test time is given. The codes are run in Matlab R2017 on a computer with Intel Core i7-9800X 3.80GHz with 16 GB RAM. The results are shown in Table VII.

It can be seen in Table VII that unsupervised methods are fastest. In these methods, images are treated as bags of pixels and only basic calculations and statistics are needed. For supervised methods, more time is required to get high dimensional image features. The test time cost of methods is close to unsupervised methods. And the time cost of the training process is also acceptable, which can be carried out offline. The test time costs of proposed methods mainly

TABLE VII
COMPUTATION TIME OF DIFFERENT METHODS

Method		Train(s)	Test(s)
Unsupervised	GE [8]	-	0.213
	GW [6]	-	0.161
	MaxRGB [4]	-	0.160
	SoG [7]	-	0.234
Supervised	SVR [20]	2022.090	3.560
	NIS [24]	607.767	1.120
	NN [19]	2049.524	3.612
Semi-supervised	SCC_L	192.721	0.347
	SCC_M	634.638	0.348
	SCC_H	413.074	0.347
	MSCC	1058.635	0.351

contain two parts: feature extraction and illumination calculation. The time cost of three SCC methods is slightly different because of different coefficient matrix. The combination step makes MSCC cost a little more time than SCC.

VI. CONCLUSION AND FUTURE WORK

Supervised methods have been shown to be superior to those unsupervised ones in illumination estimation. However, they must face a basic issue that is lack of sufficient labeled data for each specific camera. This paper presents a novel semi-supervised method to leverage both labeled and unlabeled samples. To the end, three graphs based on three visual cues, low-level RGB color distribution, mid-level initial illuminant estimates and high-level scene content, are constructed to represent the relationship among different images. Then a multi-cue semi-supervised color constancy method (MSCC) is proposed by integrating these three graphs into a unified model. Extensive evaluations on several benchmark datasets demonstrate that the proposed method outperforms the supervised approaches illumination estimation with inadequate labeled samples.

Among the three cues, methods based on high-level cues perform slightly worse while the number of training samples drops down. And the combination results may be slightly influenced in several cases due to the instable performance. Beside high-level cues we use in this work, other cues such as 3D scene geometry [26] or color moment-based feature [23] can also be exploited. In future work, the choose of optional high-level cues can be further explored. Besides, due to the different sensors of cameras, only unlabeled images which are captured under a certain type of camera can be used in our semi-supervised methods. This limitation can be another avenue of future research.

REFERENCES

[1] K. Barnard, V. Cardei, and B. Funt, "A comparison of computational color constancy algorithms. I: Methodology and experiments with synthesized data," *IEEE Trans. Image Process.*, vol. 11, no. 9, pp. 972–984, Sep. 2002.

[2] K. Barnard, L. Martin, A. Coath, and B. Funt, "A comparison of computational color constancy Algorithms. II. Experiments with image data," *IEEE Trans. Image Process.*, vol. 11, no. 9, pp. 985–996, Sep. 2002.

[3] J. Von Kries, "Influence of adaptation on the effects produced by luminous stimuli," *Sources Color Vis.*, pp. 145–148, 1970.

[4] E. H. Land, "The retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, p. 108, Dec. 1977.

[5] L. Shi and B. Funt, "MaxRGB reconsidered," *J. Imag. Sci. Technol.*, vol. 56, no. 2, pp. 1–10, Mar. 2012.

[6] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Inst.*, vol. 310, no. 1, pp. 1–26, Jul. 1980.

[7] G. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Proc. Color Imag. Conf.*, no. 1, 2004, pp. 37–41.

[8] J. V. Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2207–2214, Sep. 2007.

[9] X.-S. Zhang, S.-B. Gao, R.-X. Li, X.-Y. Du, C.-Y. Li, and Y.-J. Li, "A retinal mechanism inspired color constancy model," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1219–1232, Mar. 2016.

[10] D. Cheng, D. K. Prasad, and M. S. Brown, "Illuminant estimation for color constancy: Why spatial-domain methods work and the role of the color distribution," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 31, no. 5, p. 1049, 2014.

[11] K.-F. Yang, S.-B. Gao, and Y.-J. Li, "Efficient illuminant estimation for color constancy using grey pixels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New York, NY, USA: IEEE, Jun. 2015, pp. 2254–2263.

[12] S. Bianco and C. Cusano, "Quasi-unsupervised color constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12212–12221.

[13] G. D. Finlayson, S. D. Hordley, and P. M. HubeL, "Color by correlation: A simple, unifying framework for color constancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1209–1221, Nov. 2001.

[14] J. Vazquez-Corral, M. Vanrell, R. Baldrich, and F. Tous, "Color constancy by category correlation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1997–2007, Apr. 2012.

[15] D. A. Forsyth, "A novel algorithm for color constancy," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 5–35, Aug. 1990.

[16] A. Gijsenij, T. Gevers, and J. van de Weijer, "Generalized gamut mapping using image derivative structures for color constancy," *Int. J. Comput. Vis.*, vol. 86, nos. 2–3, pp. 127–139, Jan. 2010, doi: 10.1007/s11263-008-0171-3.

[17] D. H. Brainard and W. T. Freeman, "Bayesian color constancy," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 14, no. 7, pp. 1393–1411, Jul. 1997. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-14-7-1393

[18] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[19] V. C. Cardei, B. Funt, and K. Barnard, "Estimating the scene illumination chromaticity by using a neural network," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 19, no. 12, pp. 2374–2386, Dec. 2002. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-19-12-2374

[20] W. Xiong and B. Funt, "Estimating illumination chromaticity via support vector regression," *J. Imag. Sci. Technol.*, vol. 50, no. 4, pp. 341–348, Jul. 2006.

[21] D. Cheng, B. Price, S. Cohen, and M. S. Brown, "Effective learning-based illuminant estimation using simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1000–1008.

[22] A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color constancy with spatio-spectral statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1509–1519, Aug. 2012.

[23] G. D. Finlayson, "Corrected-moment illuminant estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1904–1911.

[24] A. Gijsenij and T. Gevers, "Color constancy using natural image statistics and scene semantics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 687–698, Apr. 2011.

[25] S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Automatic color constancy algorithm selection and combination," *Pattern Recognit.*, vol. 43, no. 3, pp. 695–705, Mar. 2010.

[26] R. Lu, A. Gijsenij, T. Gevers, V. Nedovic, D. Xu, and J.-M. Geusebroek, "Color constancy using 3D scene geometry," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1749–1756.

[27] H. R. V. Joze and M. S. Drew, "Exemplar-based color constancy and multiple illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 860–873, May 2014.

[28] S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Improving color constancy using indoor–outdoor image classification," *IEEE Trans. Image Process.*, vol. 17, no. 12, pp. 2381–2392, Dec. 2008.

[29] J. van de Weijer, C. Schmid, and J. Verbeek, "Using high-level visual information for color constancy," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[30] S. Bianco and R. Schettini, "Color constancy using faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 65–72.

[31] S. Bianco, C. Cusano, and R. Schettini, "Color constancy using CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 81–89.

[32] S. Bianco, C. Cusano, and R. Schettini, "Single and multiple illuminant estimation using convolutional neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4347–4362, Sep. 2017.

[33] Y. Hu, B. Wang, and S. Lin, "FC^4: Fully convolutional color constancy with confidence-weighted pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4085–4094.

[34] M. Buzzelli, J. van de Weijer, and R. Schettini, "Learning illuminant estimation from object recognition," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3234–3238.

[35] S. W. Oh and S. J. Kim, "Approaching the computational color constancy as a classification problem through deep learning," *Pattern Recognit.*, vol. 61, pp. 405–416, Jan. 2017.

[36] W. Shi, C. C. Loy, and X. Tang, "Deep specialized network for illuminant estimation," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 371–387.

[37] J. T. Barron, "Convolutional color constancy," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 379–387.

[38] J. T. Barron and Y.-T. Tsai, "Fast Fourier color constancy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 886–894.

[39] B. Li, W. Xiong, W. Hu, B. Funt, and J. Xing, "Multi-cue illumination estimation via a tree-structured group joint sparse representation," *Int. J. Comput. Vis.*, vol. 117, no. 1, pp. 21–47, Mar. 2016.

[40] F. Ciurea and B. V. Funt, "A large image database for color constancy research," in *Proc. Color Imag. Conf.*, 2003, pp. 160–164.

[41] B. Li, W. Xiong, W. Hu, and B. Funt, "Evaluating combinational illumination estimation methods on real-world images," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1194–1209, Mar. 2014.

[42] W. Fei, W. Xin, and L. Tao, "Semi-supervised multi-task learning with task regularizations," in *Proc. 9th IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 562–568.

[43] B. Kågström and P. Poromaa, "LAPACK-style algorithms and software for solving the generalized Sylvester equation and estimating the separation between regular matrix pairs," *ACM Trans. Math. Softw.*, vol. 22, no. 1, pp. 78–103, Mar. 1996.

[44] L. Shi and B. Funt. (2011). *Re-Processed Version of the Gehler Color Constancy Dataset of 568 Images*. [Online]. Available: http://www.cs.sfu.ca/~colour/data/

[45] [Online]. Available: http://www.colorconstancy.com/

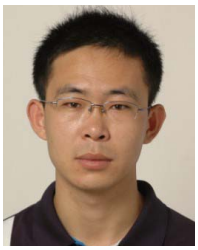
[46] G. Finlayson, G. Hemrit, A. Gijsenij, and P. Gehler, "A curious problem with using the colour checker dataset for illuminant estimation," in *Proc. 25th Color Imag. Conf. Final Program*. Springfield, VA, USA: Society for Imaging Science and Technology, Sep. 2017, pp. 64–69.

[47] G. Hemrit *et al.*, "Rehabilitating the colorchecker dataset for illuminant estimation," in *Proc. 26th Color Imag. Conf. (CIC)*. Springfield, VA, USA: Society for Imaging Science and Technology, Nov. 2018, pp. 350–353.

[48] G. Hemrit *et al.*, "Providing a single ground-truth for illuminant estimation for the ColorChecker dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1286–1287, May 2020.

Xinwei Huang received the M.S. degree from Beijing Jiaotong University, China, in 2018. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests include image processing, color constancy, and 3D shape matching.

Bing Li received the Ph.D. degree from the Department of Computer Science and Engineering, Beijing Jiaotong University, Beijing, China, in 2009. From 2009 to 2011, he worked as a Postdoctoral Research Fellow with the National Laboratory of Pattern Recognition (NLPR). He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include computer vision, color constancy, visual saliency detection, multi-instance learning, and data mining.



Shuai Li (Member, IEEE) received the Ph.D. degree in computer science from Beihang University. He is currently an Associate Professor with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, and the Beihang Qingdao Research Institute. His research interests include computer graphics, pattern recognition, computer vision, and medical image processing.

Wenjuan Li received the Ph.D. degree from Tianjin University, China, in 2017. She is currently a Postdoctoral Research Fellow with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests include image processing, color constancy, and pattern recognition.

Weihua Xiong received the Ph.D. degree from the Department of Computer Science, Simon Fraser University, Canada, in 2007. His research interests include color science, computer vision, color image processing, and stereo vision.

Xuanwu Yin received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2011 and 2017, respectively. Since 2017, he has been with the Department of Kirin Chipset and Technology Development, Hisilicon, focusing on image signal processing algorithms. His current research interests include color science, computational color constancy, and color reproduction.

Weiming Hu (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University, in 1998. He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences.

His research interests include visual surveillance and filtering of Internet objectionable information.

Hong Qin received the B.S. and M.S. degrees in computer science from Peking University, and the Ph.D. degree in computer science from the University of Toronto. He is currently a Professor of computer science with the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing.