frontiers | Frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 08 April 2022
doi: 10.3389/fnbot.2022.841426

Color Constancy *via* Multi-Scale Region-Weighed Network Guided by Semantics

Fei Wang[1,2]*, *Wei Wang*[3], *Dan Wu*[1] and *Guowang Gao*[1]

[1] School of Electronic Engineering, Xi'an Shiyou University, Xi'an, China, [2] State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, Changsha, China, [3] School of Telecommunications Engineering, Xidian University, Xi'an, China

In obtaining color constancy, estimating the illumination of a scene is the most important task. However, due to unknown light sources and the influence of the external imaging environment, the estimated illumination is prone to color ambiguity. In this article, a learning-based multi-scale region-weighed network guided by semantic features is proposed to estimate the illuminated color of the light source in a scene. Cued by the human brain's processing of color constancy, we use image semantics and scale information to guide the process of illumination estimation. First, we put the image and its semantics into the network, and then obtain the region weights of the image at different scales. After that, through a special weight-pooling layer (WPL), the illumination on each scale is estimated. The final illumination is calculated by weighting each scale. The results of extensive experiments on Color Checker and NUS 8-Camera datasets show that the proposed approach is superior to the current state-of-the-art methods in both efficiency and effectiveness.

Keywords: color constancy, multi-scale, weight pooling layer, semantic, network

OPEN ACCESS

Edited by:
Kok-Lim Alvin Yau,
Tunku Abdul Rahman University, Malaysia

Reviewed by:
Masataka Sawayama,
Institut National de Recherche en Informatique et en Automatique (INRIA), France
Shaobing Gao,
University of Electronic Science and Technology of China, China
Fengshun Lu,
Beijing Aerohydrodynamic Frontier Research Center, China
Kannimuthu Subramanian,
Karpagam Academy of Higher Education, India

***Correspondence:**
Fei Wang
200102@xsyu.edu.cn

***Received:** 22 December 2021*
***Accepted:** 08 March 2022*
***Published:** 08 April 2022*

Citation:
Wang F, Wang W, Wu D and Gao G (2022) Color Constancy via Multi-Scale Region-Weighed Network Guided by Semantics. Front. Neurorobot. 16:841426. doi: 10.3389/fnbot.2022.841426

1. INTRODUCTION

The observed color of an object in an image (representing the observed values in RGB space) depends on the intrinsic color and light-source color. It is quite easy to distinguish the reflectance from the light-source color for human beings while endowing a computer with the same ability is difficult (Gilchrist, 2006). For example, given a red object, how can one discern if it is a white object under red light or a red object under a white light? To assist a computer in solving this problem, it is necessary to separate the color of the light source, namely, the color constancy.[1] The goal of computational color constancy is to preserve the perceptive colors of objects under different lighting conditions by removing the effect of color casts caused by the scene's illumination.

Color constancy is a fundamental research topic in the image-processing and computer-vision fields, and it has many applications in photographic technology, object recognition, object detection, image segmentation, and other version systems. Color casts caused by incorrectly applied computational color constancy can negatively impact the performance of image segmentation and classification (Afifi and Brown, 2019; Xue et al., 2021), thus, there is a rich body of work on this

[1] In this study, we aim to solve the color-constancy problem with a single light source.

topic. Generally, methods for obtaining color constancy with image data are divided into two main categories: low-level-feature-based methods (Buchsbaum, 1980; Brainard and Wandell, 1986; Lee, 1986; Wandell and Tominaga, 1989; Nieves et al., 2000; Krasilnikov et al., 2002; Weijer et al., 2007; Gehler et al., 2008; Tan et al., 2008; Toro, 2008; Gijsenij et al., 2011; Finlayson, 2013; Gao et al., 2013; Barron, 2015; Bianco et al., 2015, 2017; Cheng et al., 2015; Shi et al., 2016; Xiao et al., 2020; Yu et al., 2020) and semantic-feature-based methods (Schroeder and Moser, 2001; Spitzer and Semo, 2002; Van De Weijer et al., 2007; Bianco et al., 2008; Lau, 2008; Li et al., 2008; Gao et al., 2015; Afifi, 2018).

Low-level-features-based methods pay attention to the law of the color of the image itself, and they do not consider the image-content information. These methods consider the relationship between color and achromatic color statistics (Weijer et al., 2007; Gehler et al., 2008), inspired by the human visual system (Nieves et al., 2000; Krasilnikov et al., 2002; Gao et al., 2013), spatial derivatives, and frequency information of scene illuminations on the image (Nayar et al., 2007; Joze and Drew, 2014), extract hand-crafted features from training data (Buchsbaum, 1980; Brainard and Wandell, 1986; Finlayson, 2013; Cheng et al., 2015), and learn features automatically by a convolutional neural network (CNN) from samples (Barron, 2015; Bianco et al., 2015, 2017; Shi et al., 2016; Xiao et al., 2020; Yu et al., 2020). Although these methods have achieved good results, especially the CNN-based methods, various methods are used to make the illumination estimation as accurate as possible, but in some complex situations, due to inflexibility, they cannot well solve the color ambiguity.

Semantic-feature-based methods are more in line with human vision. When observing a scene, human beings have a certain psychological memory of the color of the object itself in the scene. Therefore, the content of the scene can play a certain guiding role in color constancy. Because previous attempts at semantic information extraction have not been accurate, there is relatively little research on this type of algorithm. Van De Weijer et al. (2007) proposed a color-constancy algorithm based on advanced visual information. The algorithm models the image into many semantic categories, such as sky, grassland, road, pedestrian, and vehicle, and it calculates multiple possible illumination values from these semantic categories. Each illumination is used to correct the image and calculate the semantic combination with the greatest probability. At this time, the illumination is the optimal scene illumination. Schroeder and Moser (2001) divided images into different categories, and then they learned different features for each type. Bianco et al. (2008) proposed an indoor and outdoor adaptive illumination estimation algorithm that uses a classification algorithm to divide the image into indoor and outdoor scenes, and then they estimated the illumination according to the parameters learned by training data. Afifi (2018) exploited the semantic information together with the color and spatial information of the input image, and they trained a CNN to estimate the illuminant color and gamma correction parameters. This is one of the most effective methods of this type.

However, these two methods may not find the optimal solution in some complex situations due to inflexibility. To summarize, several open problems remain unsolved in these approaches, which can be generally concluded to have two aspects.

- *Color ambiguity with only low-level features:* Many of these methods (Buchsbaum, 1980; Brainard and Wandell, 1986; Lee, 1986; Wandell and Tominaga, 1989; Nieves et al., 2000; Krasilnikov et al., 2002; Weijer et al., 2007; Gehler et al., 2008; Tan et al., 2008; Toro, 2008; Gijsenij et al., 2011; Finlayson, 2013; Gao et al., 2013; Cheng et al., 2015) only focus on the color of the image itself, for images with large color deviation, it is difficult to accurately estimate the illumination. The development of CNNs has facilitated a qualitative leap in illumination estimation, but many CNN-based methods (Barron, 2015; Shi et al., 2016; Bianco et al., 2017) are patches-based, which take the small sampled image patches as input and learn the corresponding local estimations subsequently pooled into a global result. The small patches contain less contextual information, which commonly leads to ambiguity in local estimation. When inferring the illumination color in a patch, it is often the case that the patch contains little or no semantic context benefiting its reflectance or illumination estimation. To solve this problem, Hu et al. (2017) used a global image as input, and they designed a confidence weight layer to learn the weight of each patch. Afifi and Brown (2020) proposed an end-to-end approach to learn the correct white balance, which consists of a single encoder and multiple decoders, mapping an input image to two additional white-balance settings corresponding to indoor and outdoor illuminations. Both methods achieved great success. However, due to the lack of attention to semantic information, large errors in illumination estimation exist in some scenes, which has also been found in our experiments.
- *Inaccurate illumination with only semantics:* Owing to the low accuracy of semantic segmentation, the early semantic-based color-constancy algorithm has great limitations. CNNs have greatly improved the accuracy of semantic segmentation. Afifi (2018) exploited the semantic information together with the color and spatial information of the input image, and they trained a CNN to estimate the illuminant color and gamma correction parameters. However, there is a possibility of error in segmentation, and incorrect segmentation will lead to errors in illumination estimation. In our experiments, we also verified that incorrect segmentation can lead to incorrect illumination estimates.

To address the aforementioned open problems, we did some experiments. In one experiment, we conducted, yellow banana and a red apple were placed into a scene, illuminated with different colors of light, and then different observers were allowed to view the results. It was found that the observers can correctly distinguish the color of the fruit because humans generally think that bananas are yellow and apples are red. It can be seen that objects with inherent colors can guide observers to estimate the lighting of the scene; i.e., the objects in the scene have a great effect on the color constancy of human vision. In addition, we conducted an experiment in which some objects were placed in the scene to allow the observer to observe them from different distances. It was found that the colors of certain areas in the

scene observed at different distances were biased (this deviation is relatively small), but the bias did not have much impact on the overall color. In addition, in the present study, traditional algorithms were used to estimate the illumination of the image at different scales. It was found that the estimated illumination at different scales exhibits some deviations, but they were all close to the actual illumination. It can be thought that scale has a certain influence on the color constancy of human vision.

Estimating multiple illuminations from one image at multiple scales is also in line with a verification conclusion obtained in Shi et al. (2016), namely, that multiple hypothetical illuminations from one image can help improve the accuracy of illumination estimation. Inspired by that, we propose herein a learning-based **M**ulti-scale **R**egion-**w**eighed **N**etwork guided by **S**emantics (MSRWNS) to estimate the illuminated color of the light source in a scene. First, the semantic context of an image is extracted, the image and its semantics are put into the network, and through a series of convolution layers, the region weights of the image at different scales are obtained. Then, through the weight-pooling layer (WPL), the illumination estimation on each scale is obtained. The global illumination is calculated by weighting on each scale.

The MSRWNS network differs from the existing methods and has three contributions, which follow.

- It estimates multiple global illuminations at different feature scales, and it obtains the final lighting by simple weighting.
- Different from previous semantic-based methods, while using semantic guidance, a new region-WPL is used. The network layer simultaneously learns the contribution and local illumination of different regions in the image at each scale. It can, thus, effectively solve the illumination estimation error caused by the semantic segmentation error.
- A large strip is used in the convolution to replace the max pooling layer in the network, which improves the speed of light estimation without reducing accuracy.

The rest of this article is organized as follows. In section 2, the structure of the proposed network and training strategy is presented, together with the related experimental content in section 3. Conclusions are given in section 4.

2. MULTI-SCALE REGION-WEIGHED NETWORK GUIDED BY SEMANTICS

Following the widely accepted simplified diagonal model (Finlayson et al., 1994; Funt and Lewis, 2000), the color of the light source is represented as

$$I_c = E_c \times R_c, c \in \{r, g, b\}, \tag{1}$$

where $I_c = \{I_r, I_g, I_b\}$ is the color image under an unknown light source, $R_c = \{R_r, R_g, R_b\}$ the color image recorded by a white-light source, and $E_c = \{E_r, E_g, E_b\}$ the light source needed to be estimated from I_c.

A new color-space model has been used by color-constancy methods (Finlayson et al., 2004; Barron, 2015; Shi et al., 2016) in recent years and has certain advantages, i.e., $log - uv$ space. [2] The calculation method proceeds as follows:

$$L_u = log(R/G), L_v = log(B/G). \tag{2}$$

After estimating the light, it can be converted back to RGB space through a very simple formula:

$$\begin{aligned} R = exp(-L_u)/z, G = 1/z, B = exp(-L_v)/z \\ z = \sqrt{\exp(-L_u)^2 + \exp(-L_v)^2 + 1}, \end{aligned} \tag{3}$$

where (L_u, L_v) is the image in $log - uv$ color space. (R, G, B) is the image in RGB color space.

2.1. Problem Formulation

Generally, we only know the image I_c under an unknown light source E_c that must be estimated. The goal of color constancy is to estimate E_c from I_c and then compute it as $E_c = I_c/R_c$. How do we estimate E_c from I_c? To address this problem, we formulate color constancy as a regression problem.

First, we obtain the semantic context of the image, and for this, we use PSPNet (Zhao et al., 2016), defined as I_s, and then convert I_c from RGB space to $log - uv$ space to obtain (I_u, I_v). Combining these three channels into a new three-channel image I_n, the aim is to find a mapping f_θ such that $f_\theta(I_n) = P_{uv}$, where P_{uv} represents the light value in $log - uv$ space.

As mentioned earlier, the objects in the scene have a great effect on the color constancy of human vision (Van De Weijer et al., 2007; Gao et al., 2019). Therefore, the designed color-constancy algorithm should imitate the human visual system, i.e., the mapping f_θ should be able to be based on semantic information and is used to support the larger contribution area and suppress the smaller contribution area in the image. Therefore, two aspects must be considered in the model: First, one must find a way to estimate the illumination of each area in the image, and, second, one must use an adaptive algorithm to integrate the illumination of these multiple areas into a global illumination. Supposing that $R = R_1, R_2, ..., R_n$ represents n non-overlapping regions in the image I_c, ${E_{uv}}^i$ represents the estimated scene illumination of the ith area R_i. Therefore, the mapping f_θ can be expressed as follows:

$$f_\theta(I_n) = P_{uv} = \sum_{i=0}^{n-1} w(R_i) E_{uv}^i, \tag{4}$$

where $w(R_i)$ represents the contribution of each area to the illumination estimation, i.e., the weight. In other words, if R_i contains the semantic context information, the corresponding $w(R_i)$ has a higher weight value.

[2] As reported in Finlayson et al. (2004), Barron (2015), $log - uv$ is better than RGB, since, first, there are two variables instead of three; and second, the multiplicative constraint of the illumination estimation model is converted to the linear constraint.

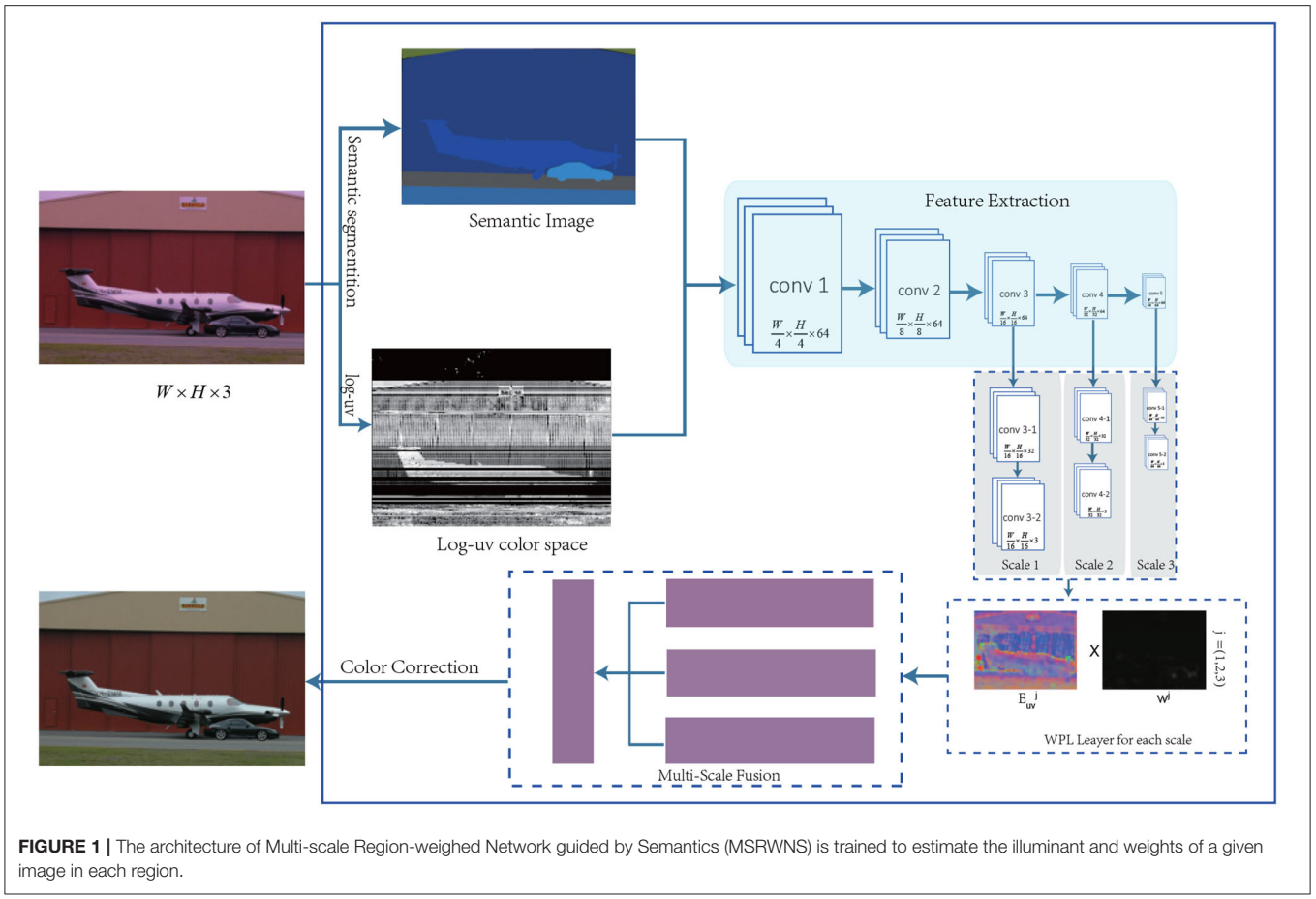

FIGURE 1 | The architecture of Multi-scale Region-weighed Network guided by Semantics (MSRWNS) is trained to estimate the illuminant and weights of a given image in each region.

2.2. Network Architecture

It can be seen from Equation (4) that it is necessary to design a network structure f_θ to be able to calculate $w(R_i)$ and E^i_{uv} in each area. The network structure is shown in **Figure 1**.

We learn P_{uv} at different scales from the intermediate features with scales of 1/16, 1/32, and 1/64. Defining the superscript j to represent the scale, ${P_{uv}}^j$ then represents the estimated illumination in $log - uv$ space under the jth scale, which is converted back to RGB space according to Equation (3) to obtain ${P_c}^j$, where $c = R, G, B$. Finally, the final illumination P_c in RGB color space is obtained by simple calculation of the obtained illumination on different scales, and the formula is:

$$\begin{aligned} P_c &= \sum_{j=1}^{n} C_j P_c^j \\ \sum_{j=1}^{n} C_j &= 1, \end{aligned} \tag{5}$$

where C_j represents the weight of the illumination obtained at each scale. In this article, $j = 1, 2, 3$. In the final illumination calculation, it is assumed that the estimated illumination at different scales contributes the same to the scene illumination, namely, $C_j = 1/3, j = 1, 2, 3$.

2.3. Weight-Pooling Layer

In most of the previous methods, the extracted features are directly calculated through several fully connected layers to obtain a global illumination. However, it can be seen from the results of the earlier literature that the effect of illumination estimation is not significantly improved. Referring to Hu et al. (2017), we used a custom network layer, called a WPL, the main function of which is to converge the regional illumination into a global illumination, and at the same time, learn the weight $w(R_i)$ of each area; R_i represents the ith area. The *WPL* on each scale is expressed as follows:

$$P^j = \sum_{i=0}^{n-1} {w_i}^j {E_i}^j, \tag{6}$$

where P^j represents the output of the WPL layer in the jth scale, ${w_i}^j$ the contribution of each region that must be learned on the jth scale, ${E_i}^j$ the illumination of the i area on the jth scale that must be learned, and n represents the number of areas. In this study, each scale is $n = \frac{W}{16} \times \frac{H}{16}, \frac{W}{32} \times \frac{H}{32}, \frac{W}{64} \times \frac{H}{64}$.

2.4. Illumination Fusion for Multiple Scales

We made a simple attempt to determine how to select illumination at multiple scales, and we used several samples

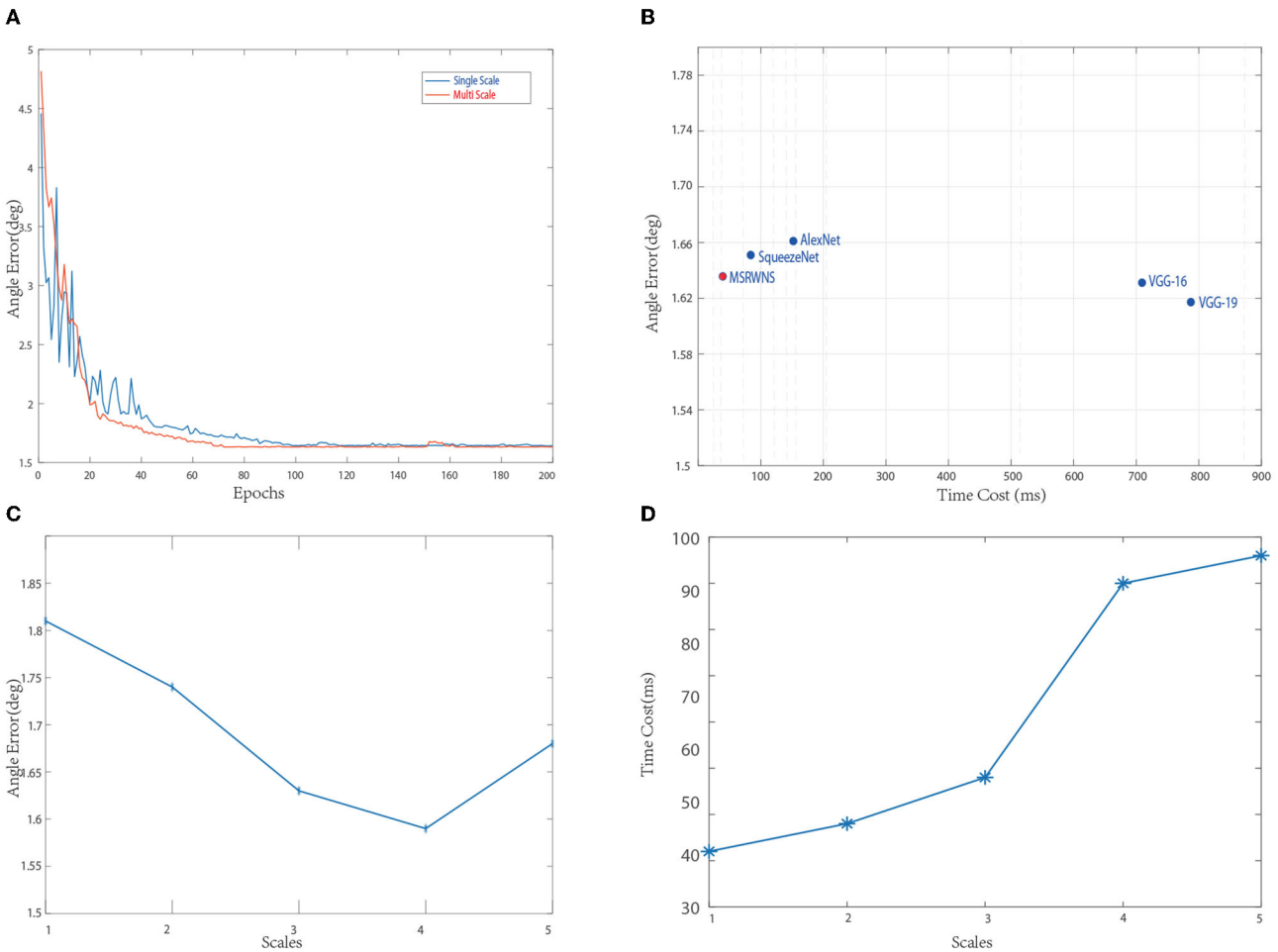

FIGURE 2 | (A–D) Performance under different parameters.

TABLE 1 | Accuracy of different network input sizes and whether semantics are used or not in the Color Checker dataset.

Size/Method	Mean	Median	TriMean	Best 25%	Worst 25%	Speed(ms)
256,T	1.67	1.36	1.53	0.45	4.09	27
256,None	1.81	1.44	1.62	0.55	4.31	23
512,T	1.64	1.17	1.28	0.31	3.82	34
512,None	1.66	1.33	1.48	0.42	4.01	30
128,T	1.74	1.33	1.49	0.55	4.17	19
128,None	1.76	1.35	1.51	0.55	4.27	15

256, 512, *and* 128 *are defined as different input sizes.* T, None *defined as the use of semantics or not. Red indicates best accuracy.*

to train the 3 classification problems, hoping to obtain the probability of illumination at different scales. However, the training process model is difficult to converge and the effect is not ideal. Finally, for the sake of simplicity, it is assumed that each scale has the same contribution to the illumination estimation, so the average value of 3-scale illumination is taken as the final illumination in this section[3], and the results also show that the average value is higher than that of a single scale on a variety of datasets.

2.5. Loss Function

At the time of training optimization, Euclidian loss is utilized for the network, defined

$$Loss = \sum_{j=1}^{n} Loss_j, \tag{7}$$

[3]In this work, the average value of multiple-scale illumination is used.

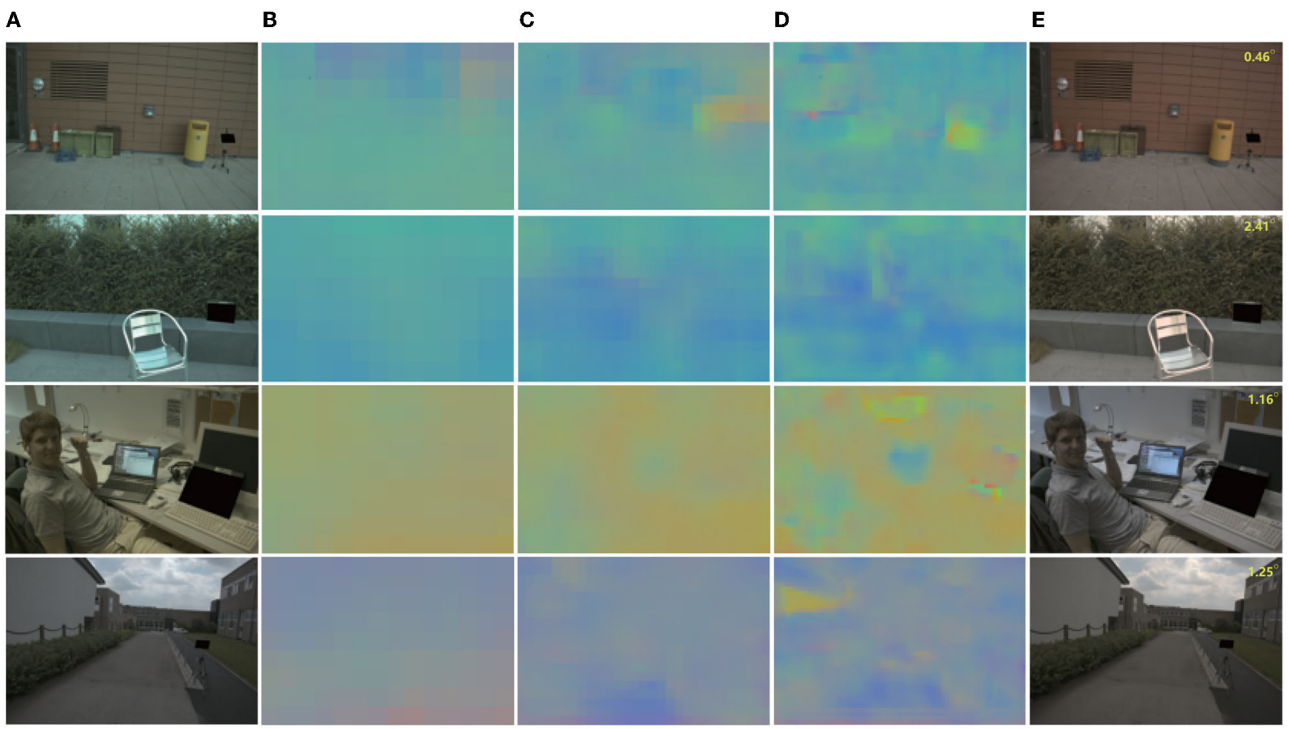

FIGURE 3 | Local illumination estimation of network output at different scales. Left to right: **(A)** original color-biased image; **(B)** Local illumination at scale 3; **(C)** Local illumination at scale 2; **(D)** Local illumination at scale 1; **(E)** Final correction results. It can be seen from the figure that under different scales, due to different area sizes, the estimated illumination is also different, but the overall color is basically the same.

$$Loss_j = \frac{1}{N}\sum_{i=1}^{n} \left\| E_i^e - E_i^t \right\|^2, \tag{8}$$

where $Loss_j$ is the loss function of the jth scale, E^e is the illumination estimated by the network, E^t is the ground truth illumination, and N represents the batch of the training samples. The loss is minimized using stochastic gradient descent with standard back-propagation.

2.6. Discussion of Network Structure

Either shallower (i.e., Shi et al. 2016) or deeper networks (i.e., VGG-net Simonyan and Zisserman 2014) could replace the pre-feature extraction in the proposed system. However, due to the color-constancy problem, the best network for feature extraction should have enough capacity to distinguish ambiguities and should be sensitive to different illuminants. We tried several common networks, such as AlexNet (Krizhevsky et al., 2012), VggNet16 (Simonyan and Zisserman, 2014), and VggNet19 (Simonyan and Zisserman, 2014), and they all achieved good results. Finally, to improve the computational efficiency of the network, we simplified AlexNet (Krizhevsky et al., 2012) and removed all of its pooling layers. The test results showed that the efficiency increased by approximately 4 and the accuracy by an average of 5.2%.

3. EXPERIMENTAL RESULTS

3.1. Datasets

Because semantic information is needed in this study, we mainly used a semantic segmentation dataset, namely, the ADE20k dataset (Zhou et al., 2016). Meanwhile, we used PSPNET (Zhao et al., 2016; Zhou et al., 2017)[4] to segment the semantic information for Color Checker (Gehler et al., 2008; Zhou et al., 2017) and NUS 8-Camera datasets (Cheng et al., 2014).

During the training process, 100 images with accurate semantic segmentation were manually selected from the Color Checker dataset (Gehler et al., 2008), and 200 images were extracted from the NUS 8-Camera dataset (Cheng et al., 2014).

In addition, since the ADE20k dataset (Zhou et al., 2017) does not provide illumination information, it is assumed that the images in the ADE20k dataset are corrected white-balanced images, and we, therefore, visually selected 500 images with normal color. Different lights were then rendered according to the following equations:

$$I_i' = I_i.M_i, \tag{9}$$

[4]We did not train the scene parsing network ourselves, the model used is downloaded from https://github.com/hszhao/PSPNet.

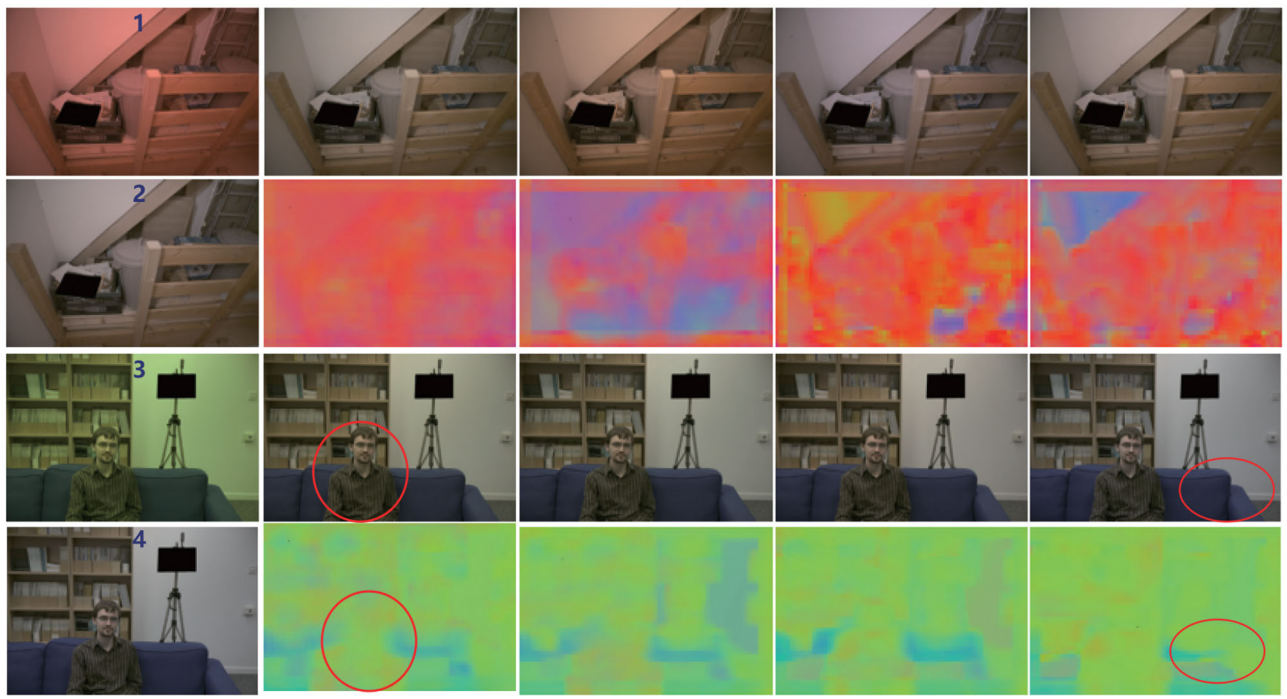

FIGURE 4 | Effect of training at different epochs. Left to right: **(A)** original image and normal image; **(B)** output image and local illumination at 100 epochs; **(C)** output image and local illumination at 500 epochs; **(D)** output image and local illumination at 2,000 epochs; **(E)** output image and local illumination 4,000 epochs. It can be seen that with an increasing number of epochs, the corrected image tends to the real result. In addition, it can be seen from the red circle that the area in which people are located and that at the junction of the wall and sofa are different in local color.

$$M_i = \begin{bmatrix} r_i & 0 & 0 \\ 0 & g_i & 0 \\ 0 & 0 & b_i \end{bmatrix}, \tag{10}$$

where (r_i, g_i, b_i) represents the simulated scene illumination. After simulation, more than 2,000 training images with illumination labels and accurate semantic information were obtained from the ADE20k (Zhou et al., 2017) dataset. These 2,000 images were cut, mirrored horizontally and vertically, and rotated from $(-30^o, 30^o)$, 90^o, 180^o, and a total of approximately 20,000 pieces of data were obtained. Similarly, the images selected from the Color Checker (Gehler et al., 2008) and NUS 8-Camera datasets (Cheng et al., 2014) were processed in the same way to obtain approximately 8,000 images. All 28,000 images were randomly cropped and normalized to 512×512 as network input. As in the previous study, 3-fold cross-validation was used for all of the datasets, and each one run was used for training, one for validation, and one for testing.

3.2. Metrics

Color-constancy algorithms are often evaluated using a distance measure, such as Euclidean distance (Land, 1977; Buchsbaum, 1980), perceptual distances (Gijsenij et al., 2009), reproduction angular error (Finlayson et al., 2016), and angular error (Hordley and Finlayson, 2004; Cheng et al., 2014; Bianco et al., 2015; Shi et al., 2016). Within these metrics, the angular error is the most widely used in this field, and most of the existing works (Cheng et al., 2014; Bianco et al., 2015; Shi et al., 2016) that tested using the Color Checker (Gehler et al., 2008), NUS 8-Camera (Cheng et al., 2014), and ADE20k datasetd (Zhou et al., 2017) reported their performance in terms of the angular error in five indexes: *Mean*, *Median*, and *TriMean* of all errors; mean of the lowest 25% of errors (*Best* 25%); and mean of the highest 25% of errors (*Worst* 25%). Hence, in the present study, we also used these indexes. In particular, although we aimed to estimate the single illuminant in this study, we also tested performance on the popular outdoor multi-illuminant dataset (Arjan et al., 2012). The angular error between the estimated illuminant E_e and ground-truth illuminant E_e^* is computed for each image as follows:

$$e = \arccos\left(\frac{E_e.E_e^*}{||E_e||.||E_e^*||}\right). \tag{11}$$

The less the value of e is the better performance of the method.

3.3. Implementation Parameters

In this subsection, the parameter settings for training our final model are given.

Feature-extraction-network selection: Different network structures, such as AlexNet (Krizhevsky et al., 2012), VGGNet-19 (Simonyan and Zisserman, 2014), and SqueezeNet (Iandola et al.,

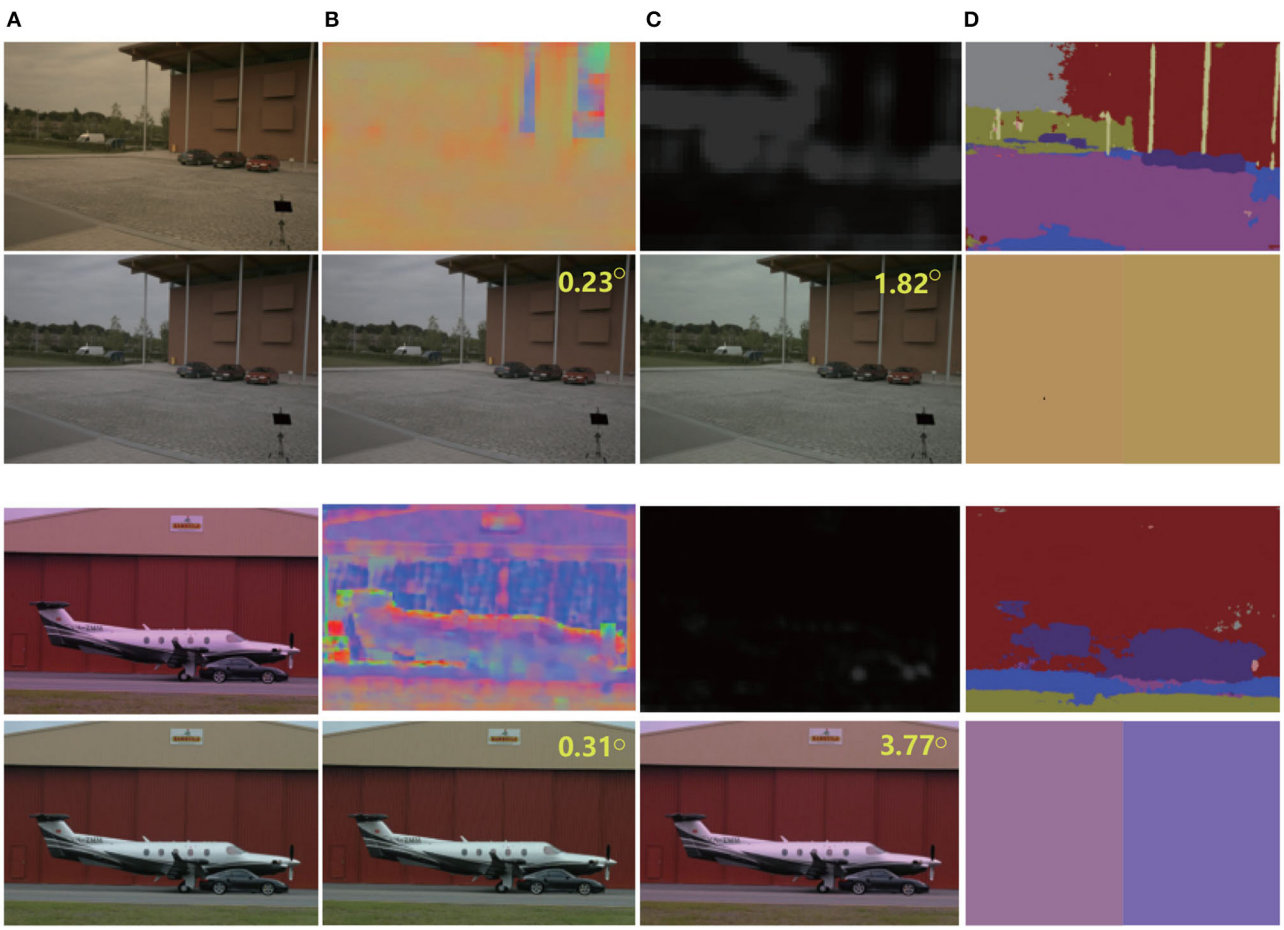

FIGURE 5 | Effect comparison between using weight-pooling layer (WPL) layer or not using it. For each group, **(A)** original image and normal image; **(B)** local illumination and image with WPL layer; **(C)** weight image and result without WPL; **(D)** semantic and global illumination (in global illumination, the left-hand side is the light estimated after using WPL and the right-hand side is that estimated without WPL). It can be seen that in an image with rich scenic content, the effect of using the WPL layer is significantly better than that of not using it. From the weighted image (the image after the weight is normalized, for which the gray value is high, the weight is large, and vice versa) and the local illumination image, different objects in the scene have different weights, and some contribute significantly. It can be seen from the estimated illumination and semantics that the illumination is also different in different semantic parts, and the approximate shape of the objects in the scene can be seen from the illumination image.

2016), were used to test performance. The comparison diagram is shown in **Figure 2B**, from which it can be seen that, although VGGNet-16 (Simonyan and Zisserman, 2014) and VGGNet-19 (Simonyan and Zisserman, 2014) network structures have a better effect than other networks, they take more time. Finally, considering effect and efficiency, the structure in **Figure 1** is used in this study.

Network input and output: We compared the performance of the model trained with (I_u, I_v, I_t) three-channel image input and (I_u, I_v) two-channel input, and we tested the performance of different resolution images as network input. The comparison results are shown in **Table 1**. Considering effect and efficiency, the effect is the best when the network input image resolution is 512 and semantic information is used at the same time.

For output, we tested the effects of 1–5 scales outputs on illumination estimation, where 1 scale uses $\frac{W}{64} \times \frac{H}{64}$, 2 scales uses $\frac{W}{32} \times \frac{H}{32}, \frac{W}{64} \times \frac{H}{64}$, 3 scales use $\frac{W}{16} \times \frac{H}{16}, \frac{W}{32} \times \frac{H}{32}, \frac{W}{64} \times \frac{H}{64}$, 4 scales use $\frac{W}{8} \times \frac{H}{8}, \frac{W}{16} \times \frac{H}{16}, \frac{W}{32} \times \frac{H}{32}, \frac{W}{64} \times \frac{H}{64}$, and 5 scales use $\frac{W}{8} \times \frac{H}{8}, \frac{W}{16} \times \frac{H}{16}, \frac{W}{32} \times \frac{H}{32}, \frac{W}{64} \times \frac{H}{64}, \frac{W}{128} \times \frac{H}{128}$. The curves are shown in **Figures 2C,D**. It can be seen that the effect is best when the scale is 4, but the time consumption is more than doubled when the scale is 3. Considering comprehensive effect and efficiency, we used 3 scales in the network. **Figure 3** shows the intermediate results of estimating illumination at different 3 scales.

Batch size and learning: For optimization, Adam (Kingma and Ba, 2014) was employed with a batch size of 64, and a basic learning rate of 0.0001 was set for training. We trained all of the experiments over 4,000 epochs (2,50,000 iterations with batch size 64). The average angular error is calculated in the Color Checker dataset for every 20 epochs. The curve is shown in **Figure 2A**. **Figure 4** shows the resulting image and local illumination after different numbers of epochs.

TABLE 2 | Performance comparison on Color Checker dataset (Gehler et al., 2008).

Method	Mean	Median	TriMean	Best 25%	Worst 25%	95th percentile
White-Patch (Brainard and Wandell, 1986)	7.55	5.68	6.35	1.45	16.12	-
Edge-based Gamut (Barnard, 2000)	6.52	5.04	5.43	1.90	13.58	-
Gray-World (Buchsbaum, 1980)	6.36	6.28	6.28	2.33	10.58	11.3
1st-order Gray-Edge (Weijer et al., 2007)	5.33	4.52	4.73	1.86	10.03	11.0
2nd-order Gray-Edge (Weijer et al., 2007)	5.13	4.44	4.62	2.11	9.26	-
Shades-of-Gray (Finlayson and Trezzi, 2004)	4.93	4.01	4.23	1.14	10.20	11.9
Bayesian (Gehler et al., 2008)	4.82	3.46	3.88	1.26	10.49	-
General Gray-World (Barnard et al., 2002)	4.66	3.48	3.81	1.00	10.09	-
Intersection-based Gamut (Gehler et al., 2008)	4.20	2.39	2.93	0.51	10.70	-
Pixel-based Gamut (Gehler et al., 2008)	4.20	2.33	2.91	0.50	10.72	14.1
Natural Image Statistics (?)	4.19	3.13	3.45	1.00	9.22	11.7
Bright Pixels (Joze et al., 2012)	3.98	2.61	-	-	-	-
Spatio-spectral (GenPrior) (Hirakawa et al., 2012)	3.59	2.96	3.10	0.95	7.61	-
Cheng et al. (2014)	3.52	2.14	2.47	0.50	8.74	-
Corrected-Moment (19 Color) (Finlayson, 2013)	3.50	2.60	-	-	-	8.6
Corrected-Moment (19 Color) (Finlayson, 2013)*	2.96	2.15	2.37	0.64	6.69	-
Corrected-Moment (19 Edge) (Finlayson, 2013)	2.82	2.00	-	-	-	6.9
Corrected-Moment (19 Edge) (Finlayson, 2013)*	3.12	2.38	2.59	0.90	6.46	-
Regression Tree (Cheng et al., 2015)	2.42	1.65	1.75	0.38	5.87	-
CNN (Bianco et al., 2015)	2.36	1.98	-	-	-	-
CCC (Barron, 2015)	1.95	1.22	1.38	0.35	4.76	5.85
DS-Net (Shi et al., 2016)	1.90	1.12	1.33	0.31	4.84	5.99
FC4 (Hu et al., 2017)	1.77	1.11	1.29	0.34	4.29	5.44
MSRWNS-AVG	1.93	1.38	1.42	0.33	4.32	4.20
MSRWNS-1	1.72	1.16	1.33	0.32	3.79	4.36
MSRWNS-2	1.68	1.13	1.28	0.31	3.84	4.44
MSRWNS-3	1.71	1.13	1.31	0.31	3.82	4.18
MSRWNS	1.64	1.13	1.28	0.31	3.78	4.07

For each metric, red indicates the best performance and blue indicates the second-best performance.

3.4. Comparison With State-of-the-Art Methods

To evaluate the performance of the proposed method and the influence of the WPL layer on the method. We trained two models. One used the mean value when the local region converges to the global, which is defined as *MSRWNS-AVG*, the other model used the *WPL* layer, which is defined as *MSRWNS*. In addition, we also estimated the illumination effect at each scale, defined as *MSRWNS-1*, *MSRWNS-2*, and *MSRWNS-3*. Several visualizations of processing outputs obtained using the proposed method are presented in **Figure 5**.

The quantitative performance comparison on the Color Checker dataset is presented in **Table 2** and the results on the NUS 8-Camera dataset in **Table 3**. The performance comparisons on the ADE20k (Zhou et al., 2017), SFU Lab dataset (Barnard et al., 2010), and SFU Gray-Ball dataset (200, 2003) are shown in **Tables 4–6**, respectively. Most CNN-based methods compare the effects on only two datasets, Color Checker dataset and NUS 8-Camera dataset. In order to make the comparison results consistent, the data in **Tables 2**, **3** are from Shi et al. (2016), while others were trained by us with the same training samples mentioned in the datasets section.

In addition, we compared our method with most of the existing works; typical works include the Deep Specialized Network (DS-Net) (Shi et al., 2016) and Fully Convolutional Color Constancy with confidence-weighted pooling (FC4) (Hu et al., 2017). Several visualizations of testing outputs obtained using the proposed method are presented in **Figures 6**, **7**.

From **Tables 2**, **3**, it can be seen that the mean error of the proposed method is reduced by 12.3% on the Color Checker dataset and by 5.8% compared to DS-Net (Shi et al., 2016), and reduced by 7.3% on the Color Checker dataset and reduced by 0.5% compared to FC4 (Hu et al., 2017). In addition, using the WPL layer under a single scale gives a better result than that obtained using the mean. The effect of using the mean of three scales on most indicators is better than the result of using a single scale.

In particular, the proposed method shows the best performance on the ADE20k (Zhou et al., 2017) dataset. The mean angular error is lower than that of DS-Net (Shi et al., 2016) by 29.7% (from 1.68 to 1.18), and the mean error of the worst 25% was reduced by 25.6% (from 3.86 to 2.87) compared to DS-Net (Shi et al., 2016). Other indicators have also been reduced to a certain extent. The reason for these results is that

TABLE 3 | Performance comparison on NUS 8-Camera dataset (Cheng et al., 2014).

Method	Mean	Median	TriMean	Best 25%	Worst 25%
White-Patch (Brainard and Wandell, 1986)	10.62	10.58	10.49	1.86	19.45
Edge-based Gamut (Barnard, 2000)	8.43	7.05	7.37	2.41	16.08
Pixel-Based Gamut (Gehler et al., 2008)	7.70	6.71	6.90	2.51	14.05
Intersection-based Gamut (Gehler et al., 2008)	7.20	5.96	6.28	2.20	13.61
Gray-World (Buchsbaum, 1980)	4.14	3.20	3.39	0.90	9.00
Bayesian (Gehler et al., 2008)	3.67	2.73	2.91	0.82	8.21
Natural Image Statistics (?)	3.71	2.60	2.84	0.79	8.47
Shades-of-Gray (Finlayson and Trezzi, 2004)	3.40	2.57	2.73	0.77	7.41
Spatio-spectral (ML) (Hirakawa et al., 2012)	3.11	2.49	2.60	0.82	6.59
2nd-order Gray-Edge (Weijer et al., 2007)	3.20	2.26	2.44	0.75	7.27
Bright Pixels (Joze et al., 2012)	3.17	2.41	2.55	0.69	7.02
1st-order Gray-Edge (Weijer et al., 2007)	3.20	2.22	2.43	0.72	7.36
Spatio-spectral (GenPrior) (Hirakawa et al., 2012)	2.96	2.33	2.47	0.80	6.18
Corrected-Moment (19 Edge)(Finlayson, 2013)*	3.03	2.11	2.25	0.68	7.08
Corrected-Moment (19 Color)(Finlayson, 2013)*	3.05	1.90	2.13	0.65	7.41
Cheng et al. (2014)	2.96	2.04	2.24	0.62	6.61
CCC (Barron, 2015)	2.38	1.48	1.69	0.45	5.85
Regression Tree (Cheng et al., 2015)	2.36	1.59	1.74	0.49	5.54
DS-Net (Shi et al., 2016)	2.24	1.46	1.68	0.48	6.08
FC4 (Hu et al., 2017)	2.12	1.53	1.67	0.48	4.78
MSRWNS-AVG	2.13	1.51	1.72	0.58	5.44
MSRWNS-1	2.12	1.45	1.64	0.46	5.12
MSRWNS-2	2.11	1.45	1.66	0.51	5.29
MSRWNS-3	2.11	1.46	1.67	0.47	5.33
MSRWNS	2.11	1.45	1.64	0.45	4.77

For each metric, red indicates the best performance and blue indicates the second-best performance.

TABLE 4 | Performance comparison on ADE20k dataset (Zhou et al., 2017).

Method	Mean	Median	TriMean	Best 25%	Worst 25%
CCC (disc+ext) (Barron, 2015)	2.14	1.66	1.82	0.32	4.24
CNN (Bianco et al., 2015)	1.96	1.32	1.14	0.23	3.94
DS-Net (Shi et al., 2016)	1.68	0.96	1.06	0.26	3.86
FC4 (Hu et al., 2017)	1.56	1.32	1.02	0.33	3.86
MSRWNS-AVG	1.66	0.96	1.44	0.32	3.66
MSRWNS	1.18	0.61	0.83	0.11	2.87

For each metric, red indicates the best performance and blue indicates the second-best performance.

TABLE 5 | Performance comparison on SFU Lab dataset (Barnard et al., 2010).

Method	Mean	Median	TriMean	Best 25%	Worst 25%
CCC (disc+ext) (Barron, 2015)	3.77	2.19	2.21	0.42	8.87
CNN (Bianco et al., 2015)	3.18	2.31	2.40	0.39	6.98
DS-Net (Shi et al., 2016)	2.93	2.11	2.23	0.34	5.87
FC4 (Hu et al., 2017)	2.99	1.78	2.11	0.33	4.62
MSRWNS-AVG	3.15	1.88	2.01	0.32	4.88
MSRWNS	2.82	1.71	1.85	0.26	4.65

For each metric, red indicates the best performance and blue the second-best performance.

TABLE 6 | Performance comparison on SFU Gray-Ball dataset (200, 2003).

Method	Mean	Median	TriMean	Best 25%	Worst 25%
CCC (disc+ext) (Barron, 2015)	3.31	1.66	1.82	0.32	4.24
CNN (Bianco et al., 2015)	2.96	1.32	1.14	0.23	3.94
DS-Net (Shi et al., 2016)	2.41	0.96	1.06	0.26	3.86
FC4 (Hu et al., 2017)	2.33	1.12	1.46	0.41	3.76
MSRWNS-AVG	2.18	1.12	1.34	0.26	3.88
MSRWNS	1.83	0.82	0.94	0.20	3.65

For each metric, red indicates the best performance and blue indicates the second-best performance.

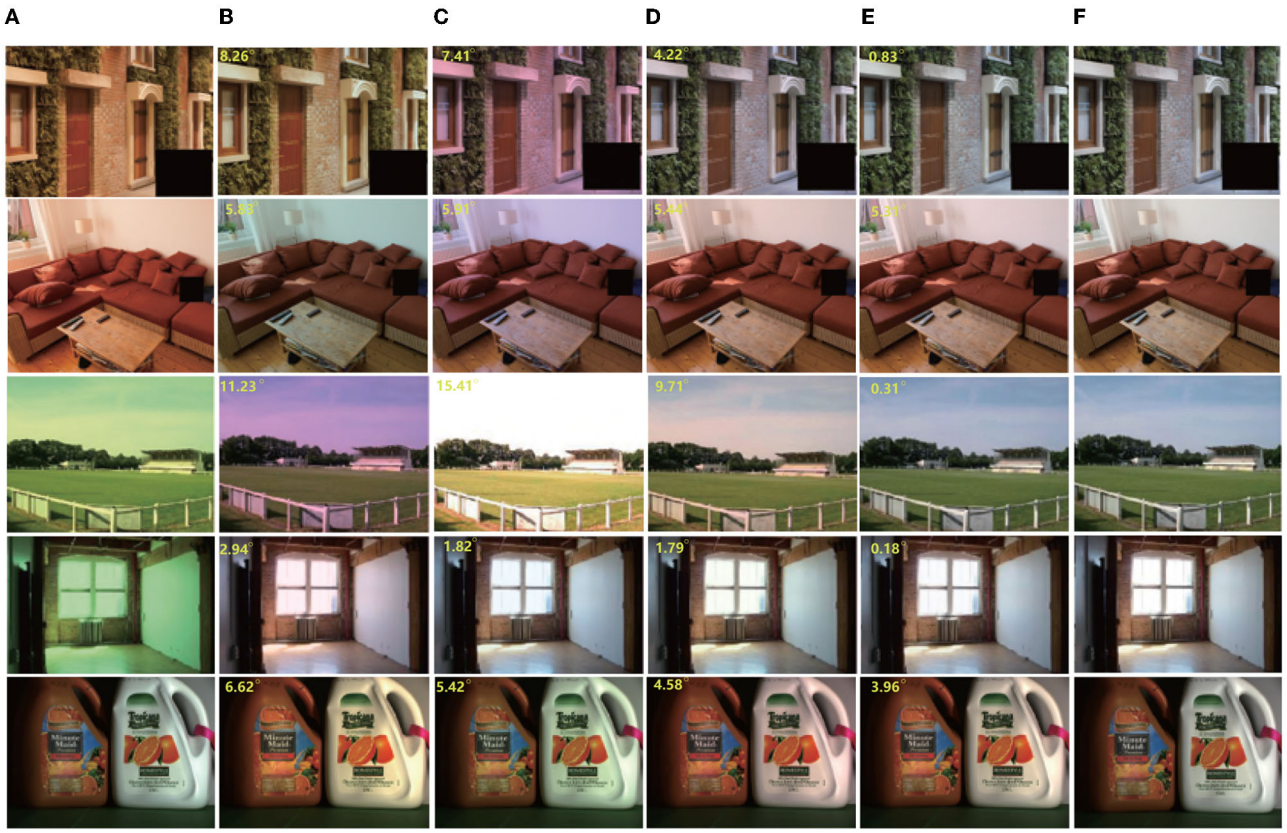

FIGURE 6 | Visual comparison results. **(A)** Original image; **(B)** result obtained by CNN (Bianco et al., 2015); **(C)** result obtained by DS-Net (Shi et al., 2016); **(D)** result obtained by FC4 (Hu et al., 2017); **(E)** result obtained by proposed method; **(F)** ground truth. Regardless of quantification or visual effects, the proposed method shows better performance, especially in the first, second, and fourth lines, because there is a large area in the scene that can be accurately segmented, and the correction result of the method proposed in this article is very close to the real image. In the second line, because the color of the sofa and that of the light are relatively close. Although the network model considers the contribution of different regions, it is difficult to eliminate the color cast caused by the similar color of the light and the surface of the object. From the corrected image look, the image has a slight red tint. In the fourth line of the image, because the objects in the scene are too singular, the red objects on the left and the white objects on the right have greater contributions and the color of the objects on the left is biased in the result.

the semantic segmentation model used is trained based on the ADE20k (Zhou et al., 2017) dataset. The accuracy of the semantics on this dataset is high, and there are parts of training data in the ADE20k dataset (Zhou et al., 2017).

In addition, we also provided several natural examples captured by a Canon D7100 without auto-white-balance [5], as shown in **Figure 7**, and we obtained several images of natural scenes with more accurate colors from the Internet, and then performed some random color casting. Results are shown in **Figure 8**.

It can be seen From **Figure 7** that on the image taken indoors (the first line in **Figure 7**) the image corrected by a CNN (Bianco et al., 2015) is obviously yellowish, and that corrected by DS-Net is slightly greenish. The white-balance effect of the camera and our result is similar and look relatively natural. In outdoor scenes,

[5]The sensor of this camera does not have a low-pass filter, and the color filter comprises several Bayer filters that overlap each other.

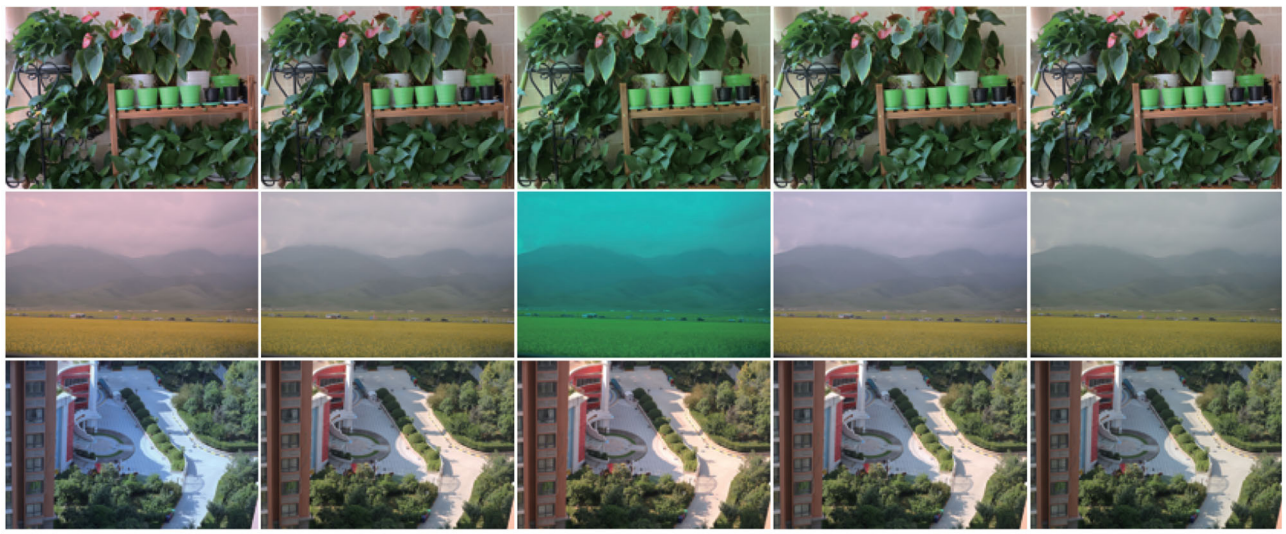

FIGURE 7 | Visual comparison between the proposed method with the auto-white-balance (AWB) function of the Canon D7100 camera used. Left to right: **(A)** original image; **(B)** results by D7100 camera with AWB; **(C)** results by CNN; **(D)** results by DS-Net; **(E)** results by present study.

it can be seen from the second line that the image corrected by a CNN (Bianco et al., 2015) has a very obvious color cast. The other types are visually more natural. Our results are seen in the sky part of the image, in which the clouds are more realistic. The scene in the third row exhibits little difference in visual effects. This may be due to sufficient sunlight in the shooting scene, and the objects in the scene receive very uniform illumination. In this way, any method can obtain better results more accurately. From **Figure 8**, it can be found from the first row that, although there is a phenomenon of highlight overflow in some areas of the sky (due to random color casts leading to pixel overflow in some areas), the overall color is close to that of the real image; the second row is due to the random color cast. The resulting color cast is small and the corrected image is basically the same as the original image. In the third row, it can be seen that, although the corrected image is different from the original image, the visual perception is better.

3.5. Efficiency

The code used to test the efficiency of the proposed method is based on Tensorflow (Rampasek and Goldenberg, 2016) and training took approximately 5 h, after which the loss tended to stabilize. In the testing phase, we converted the model to that of Caffe (Jia et al., 2014), implemented the WPL layer with C++ under Caffe, and finally used C++ code for testing. An average image took 34 ms on a CPU and only 12 ms on a GPU (the time does not include semantic segmentation) [6].

3.6. Adaptation for Multi-Illuminant

As mentioned in this article, the proposed method aims to solve the color constancy under a single illuminant, and we only compare our algorithm with existing single illuminant based methods. In addition, after the WPL layer, we can get the local illumination of the regions, it can estimate the multi-illuminant sources in different local regions, shown in **Figure 9**, where the images are taken from the popular outdoor multi-illuminant dataset (Arjan et al., 2012). However, there has a large deviation between the estimated illumination and the real multi illumination, we have analyzed the reasons and found that there are big errors in semantics. We will solve this problem in future research.

[6]experimental hardware platform: i7 7700k, 32 GB memory, gtx1080ti. If the test-image resolution was greater than 512 × 512, it was resized to 512 × 512

4. CONCLUSION

In this article, we proposed a learning based multi-scale region weighed network guided by semantics (MSRWNS) to estimate the illuminated color of the light source in a scene. Cued by the human brain's processing of color constancy, we used image semantics and scale information to guide the process of illumination estimation. First, we put an image and its semantic mask into the network and, through a series of convolution layers, the region weights of the image at different scales were obtained. Then, through a WPL, the illumination estimation on each scale was obtained. Finally, we obtained the best estimation using the weighting on each scale, and state-of-the-art performance was achieved on three of the largest color-constancy datasets, i.e., the Color Checker, NUS 8-Camera, and ADE20k datasets. This study should prove applicable in the exploration of multi-scale and semantically directed networks for other fusion tasks in computer vision. In this study, we aim to solve the color-constancy problem with a single light source, however, there are multiple light sources in the real world, in our future research, we will try to solve the problem of multiple illuminations. In addition, it is time-consuming to obtain semantics, in our future work, we will try to use semantic information only in the training phase, not in the illumination estimation phase.

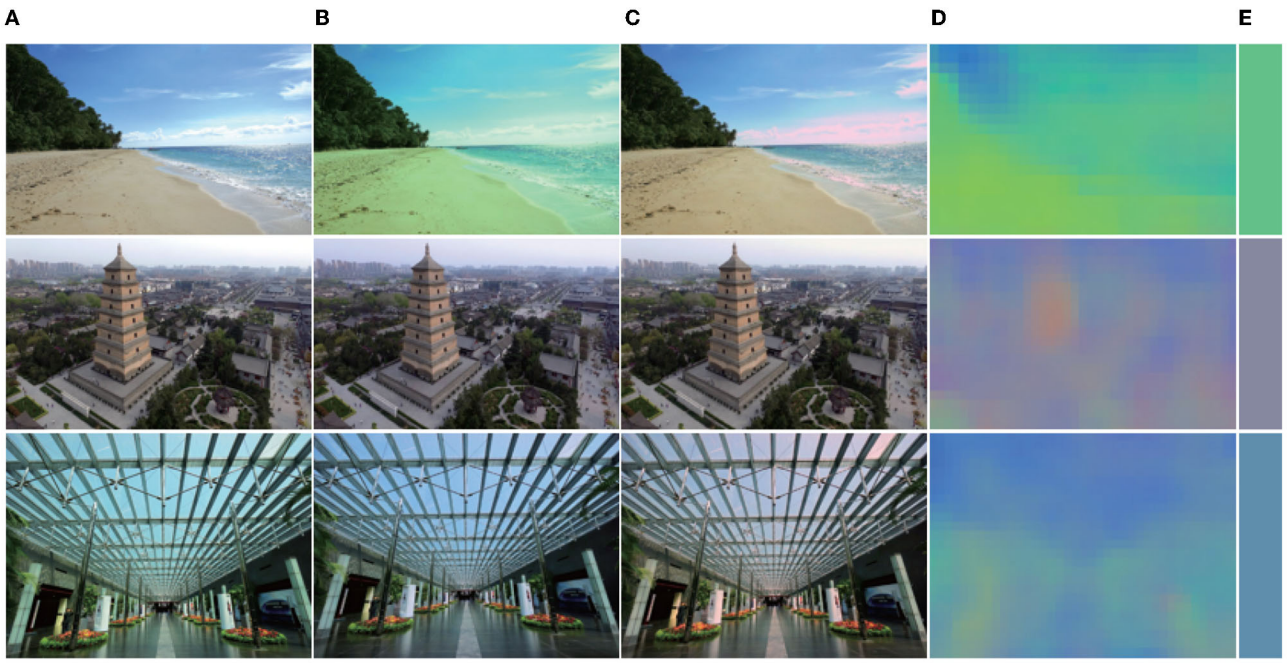

FIGURE 8 | Result in a natural scene. From left to right. **(A)** Original image; **(B)** random color-cast image; **(C)** corrected image; **(D)** local-light image; **(E)** final-light color. It can be seen from the first row that, although there is a phenomenon of highlight overflow in some areas of the sky (due to random color casts leading to pixel overflow in some areas), the overall color is close to the original image; the second row is due to the random color cast. The resulting color cast is small, and the corrected image is basically the same as the original image. In the third row, it can be seen that, although the corrected image is different from the original image, the visual perception is more realistic.

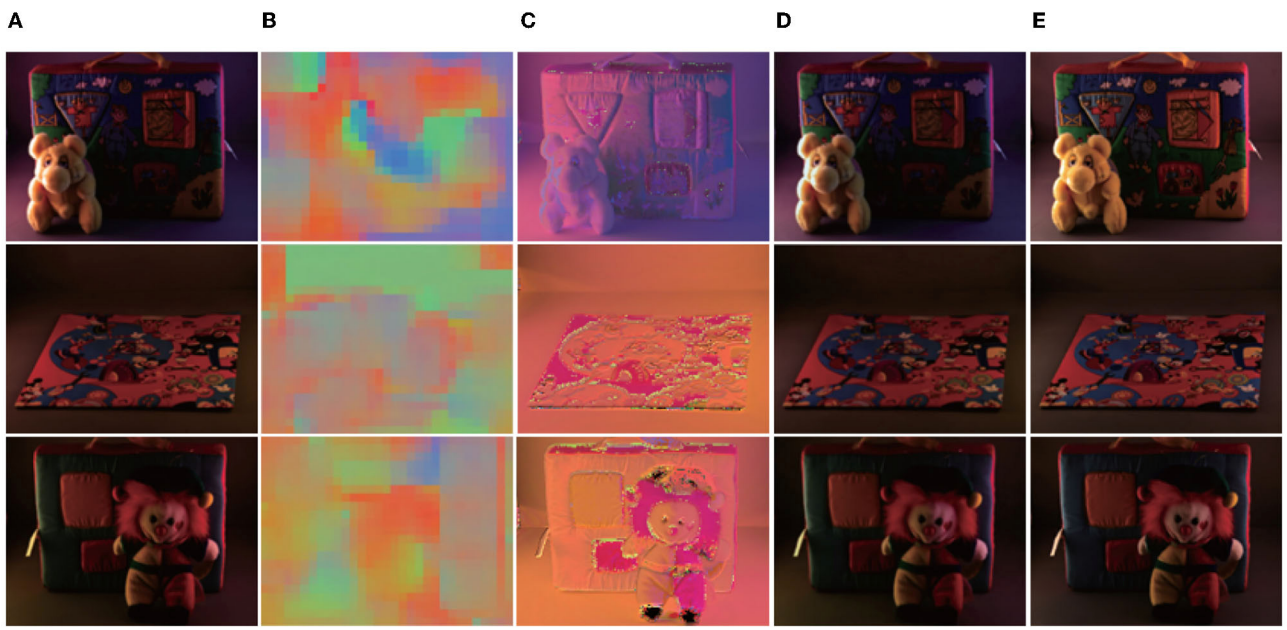

FIGURE 9 | The evolution of illumination estimated. Left to right: **(A)** original image; **(B)** the estimated local illumination; **(C)** ground-truth illumination; **(D)** the corrected image by our method; **(E)** the ground-truth image.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

FW is responsible for conceptualization, investigation, data curation, and writing. WW is responsible for formal analysis, investigation, and methodology. DW is responsible for formal analysis, investigation, and validation. GG is responsible for data curation and investigation. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by the Science Fund of State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body (No. 32015013).

REFERENCES

(2003). "A large image database for color constancy research," in *Color and Imaging Conference* (Scottsdale, AZ).

Afifi, M. (2018). Semantic white balance: Semantic color constancy using convolutional neural network. *arXiv [Preprint]*. arXiv: 1802.00153. Available online at: https://arxiv.org/pdf/1802.00153.pdf

Afifi, M., and Brown, M. S. (2019). "What else can fool deep learning? addressing color constancy errors on deep neural network performance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 243–252.

Afifi, M., and Brown, M. S. (2020). "Deep white-balance editing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Seattle, WA).

Arjan, G., Rui, L., and Theo, G. (2012). Color constancy for multiple light sources. *IEEE Trans. Image Process.* 21, 697. doi: 10.1109/TIP.2011.2165219

Barnard, K. (2000). "Improvements to gamut mapping colour constancy algorithms," in *Proc. European Conference on Computer Vision* (Dublin), 390–403.

Barnard, K., Martin, L., Coath, A., and Funt, B. V. (2002). A comparison of computational color constancy algorithms. II. experiments with image data. *IEEE Trans. Image Process.* 11, 985–996. doi: 10.1109/TIP.2002.802529

Barnard, K., Martin, L., Funt, B., and Coath, A. (2010). A data set for color research. *Color Res. Appl.* 27, 148–152.

Barron, J. T. (2015). "Convolutional color constancy," in *Proc. IEEE International Conference on Computer Vision* (Santiago), 379–387.

Bianco, S., Ciocca, G., Cusano, C., and Schettini, R. (2008). Improving color constancy using indoor-outdoor image classification. *IEEE Trans. Image Process.* 17, 2381–2392. doi: 10.1109/TIP.2008.2006661

Bianco, S., Cusano, C., and Schettini, R. (2015). Color constancy using CNNs. *arXiv [Preprint]*. arXiv: 1504. 04548. Available online at: https://arxiv.org/pdf/1504.04548.pdf

Bianco, S., Cusano, C., and Schettini, R. (2017). Single and multiple illuminant estimation using convolutional neural networks. *IEEE Trans. Image Process.* 26, 4347. doi: 10.1109/TIP.2017.2713044

Brainard, D. H., and Wandell, B. A. (1986). Analysis of the retinex theory of color vision. *J. Opt. Soc. America Opt. Image Sci.* 3, 1651.

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *J. Frankl. Inst.* 310, 1–26.

Cheng, D., Prasad, D. K., and Brown, M. S. (2014). Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *J. Opt. Soc. America Opt. Image Sci. Vis.* 31, 1049. doi: 10.1364/JOSAA.31.001049

Cheng, D., Price, B., Cohen, S., and Brown, M. S. (2015). "Effective learning-based illuminant estimation using simple features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1000–1008.

Finlayson, G. D. (2013). "Corrected-moment illuminant estimation," in *Proceedings of the IEEE International Conference on Computer Vision* (Sydney, NSW), 1904–1911.

Finlayson, G. D., Drew, M. S., and Funt, B. V. (1994). Spectral sharpening: sensor transformations for improved color constancy. *J. Opt. Soc. America Opt. Image Sci. Vis.* 11, 1553–63. doi: 10.1364/josaa.11.001553

Finlayson, G. D., Drew, M. S., and Lu, C. (2004). *Intrinsic Images by Entropy Minimization*. Heidelberg: Springer.

Finlayson, G. D., and Trezzi, E. (2004). "Shades of gray and colour constancy," in *Proc. Color and Imaging Conference* (Scottsdale, AZ), 37–41.

Finlayson, G. D., Zakizadeh, R., and Gijsenij, A. (2016). The reproduction angular error for evaluating the performance of illuminant estimation algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1482–1488. doi: 10.1109/TPAMI.2016.2582171

Funt, B. V., and Lewis, B. C. (2000). Diagonal versus affine transformations for color correction. *J. Opt. Soc. America Opt. Image Sci. Vis.* 17, 2108–2112. doi: 10.1364/josaa.17.002108

Gao, S., Yang, K., Li, C., and Li, Y. (2013). "A color constancy model with double-opponency mechanisms," in *Proceedings of the IEEE International Conference on Computer Vision* (Sydney, NSW), 929–936.

Gao, S.-B., Ren, Y.-Z., Zhang, M., and Li, Y.-J. (2019). Combining bottom-up and top-down visual mechanisms for color constancy under varying illumination. *IEEE Trans. Image Process.* 28, 4387–4400. doi: 10.1109/TIP.2019.2908783

Gao, S. B., Yang, K. F., Li, C. Y., and Li, Y. J. (2015). Color constancy using double-opponency. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1973–1985. doi: 10.1109/TPAMI.2015.2396053

Gehler, P. V., Rother, C., Blake, A., Minka, T., and Sharp, T. (2008). "Bayesian color constancy revisited," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Gijsenij, A., Gevers, T., and Lucassen, M. P. (2009). Perceptual analysis of distance measures for color constancy algorithms. *J. Opt. Soc. America Opt. Image Sci. Vis.* 26, 2243. doi: 10.1364/JOSAA.26.002243

Gijsenij, A., Gevers, T., and Van, d. W. J. (2011). Computational color constancy: survey and experiments. *IEEE Trans. Image Process.* 20, 2475–2489. doi: 10.1109/TIP.2011.2118224

Gilchrist, A. (2006). *Seeing Black and White* Oxford University Press.

Hirakawa, K., Chakrabarti, A., and Zickler, T. (2012). Color constancy with spatio-spectral statistics. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1509–1519. doi: 10.1109/TPAMI.2011.252

Hordley, S. D., and Finlayson, G. D. (2004). "Re-evaluating colour constancy algorithms," in *Proceedings of the 17th International Conference on Pattern Recognition*, Vol. 1 (IEEE), 76–79.

Hu, Y., Wang, B., and Lin, S. (2017). "Fc4: fully convolutional color constancy with confidence-weighted pooling," in *IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 330–339.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint* arXiv:1602.07360.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia* (New York, NY), 675–678.

Joze, H. R. V. and Drew, M. S. (2014). "White patch gamut mapping colour constancy," in *Proc. IEEE International Conference on Image Processing* (Orlando, FL), 801–804.

Joze, H. R. V., Drew, M. S., Finlayson, G. D., and Rey, P. A. T. (2012). "The role of bright pixels in 416 illumination estimation," in *Color and Imaging Conference, Vol. 2012* (Society for Imaging Science and Technology), 41–46.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv [Preprint]*. arXiv:1412.6980. Available online at: https://arxiv.org/pdf/1412.6980.pdf

Krasilnikov, N. N., Krasilnikova, O. I., and Shelepin, Y. E. (2002). Mathematical model of the color constancy of the human visual system. *J. Opt. Technol. C Opticheskii Zhurnal* 69, 102–107. doi: 10.1364/JOT.69.000327

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 25.

Land, E. H. (1977). The retinex theory of color vision. *Sci. Am.* 237, 108.

Lau, H. Y. (2008). *Neural inspired color constancy model based on double opponent neurons.* Hong Kong University of Science and Technology, Hong Kong, China.

Lee, H. C. (1986). Method for computing the scene-illuminant chromaticity from specular highlights. *J. Opt. Soc. America Opt. Image Sci. Vis.* 3, 1694–1699.

Li, B., Xu, D., Wang, J. H., and Lu, R. (2008). "Color constancy based on image similarity," in *Transactions on Information and Systems E91-D*, 375–378.

Nayar, S. K., Krishnan, G., Grossberg, M. D., and Raskar, R. (2007). *Method for separating direct and global illumination in a scene. US Patent App. 11/624,016.* Washington, DC: U.S. Patent and Trademark Office.

Nieves, J. L., García-Beltrán, A., and Romero, J. (2000). Response of the human visual system to variable illuminant conditions: an analysis of opponent-colour mechanisms in colour constancy. *Ophthalmic Physiol. Opt. J. Brit. Coll. Ophthalmic Opticians* 20, 44. doi: 10.1046/j.1475-1313.2000.00471.x

Rampasek, L., and Goldenberg, A. (2016). Tensorflow: biology's gateway to deep learning? *Cell Syst.* 2, 12–14. doi: 10.1016/j.cels.2016.01.009

Schroeder, M., and Moser, S. (2001). "Automatic color correction based on generic content based image analysis," in *Color and Imaging Conference* (Scottsdale, AZ), 41–45.

Shi, W., Chen, C. L., and Tang, X. (2016). "Deep specialized network for illuminant estimation," in *Proc. European Conference on Computer Vision* (Amsterdam), 371–387.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv [preprint]*. arXiv:1409.1556. Available online at: https://arxiv.org/pdf/1409.1556.pdf

Spitzer, H., and Semo, S. (2002). Color constancy: a biological model and its application for still and video images. *Pattern Recognit.* 35, 1645–1659. doi: 10.1016/S0031-3203(01)00160-1

Tan, R. T., Nishino, K., and Ikeuchi, K. (2008). *Color Constancy Through Inverse-Intensity Chromaticity Space.* Boston, MA: Springer US.

Toro, J. (2008). Dichromatic illumination estimation without pre-segmentation. *Pattern Recognit. Lett.* 29, 871–877. doi: 10.1016/j.patrec.2008.01.004

Van De Weijer, J., Schmid, C., and Verbeek, J. (2007). "Using high-level visual information for color constancy," in *2007 IEEE 11th International Conference on Computer Vision* (Rio de Janeiro: IEEE), 1–8.

Wandell, B. A., and Tominaga, S. (1989). Standard surface-reflectance model and illuminant estimation. *J. Opt. Soc. America A* 6, 576–584.

Weijer, J. V. D., Gevers, T., and Gijsenij, A. (2007). Edge-based color constancy. *IEEE Trans. Image Process.* 16, 2207–2214. doi: 10.1109/TIP.2007.901808

Xiao, J., Gu, S., and Zhang, L. (2020). "Multi-domain learning for accurate and few-shot color constancy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 3258–3267.

Xue, S., Gao, S., Tan, M., He, Z., and He, L. (2021). "How does color constancy affect target recognition and instance segmentation?" in *Proceedings of the 29th ACM International Conference on Multimedia* (New York, NY), 5537–5545.

Yu, H., Chen, K., Wang, K., Qian, Y., and Jia, K. (2020). "Cascading convolutional color constancy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34 (New York, NY), 12725–12732.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2016). "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2881–2890.

Zhou, B., Hang, Z., Fernandez, F. X. P., Fidler, S., and Torralba, A. (2017). "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 633–641.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2016). Semantic understanding of scenes through the ade20k dataset.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.