Transfer Learning for Color Constancy via Statistic Perspective

Yuxiang Tang, Xuejing Kang, Chunxiao Li, Zhaowen Lin, Anlong Ming*

School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications
{tangyuxiang, kangxuejing, chunxiaol, linzw, mal}@bupt.edu.cn

Abstract

Color Constancy aims to correct image color casts caused by scene illumination. Recently, although the deep learning approaches have remarkably improved on single-camera data, these models still suffer from the seriously insufficient data problem, resulting in shallow model capacity and degradation in multi-camera settings. In this paper, to alleviate this problem, we present a Transfer Learning Color Constancy (TLCC) method that leverages cross-camera RAW data and massive unlabeled sRGB data to support training. Specifically, TLCC consists of the Statistic Estimation Scheme (SE-Scheme) and Color-Guided Adaption Branch (CGA-Branch). SE-Scheme builds a statistic perspective to map the camera-related illumination labels into camera-agnostic form and produce pseudo labels for sRGB data, which greatly expands data for joint training. CGA-Branch further promotes efficient transfer learning from sRGB to RAW data by extracting color information to regularize the backbone's features adaptively. Experimental results show the TLCC has overcome the data limitation and model degradation, outperforming the state-of-the-art performance on popular benchmarks. Moreover, the experiments also prove the TLCC is capable of learning new scenes information from sRGB data to improve accuracy on the RAW images with similar scenes.

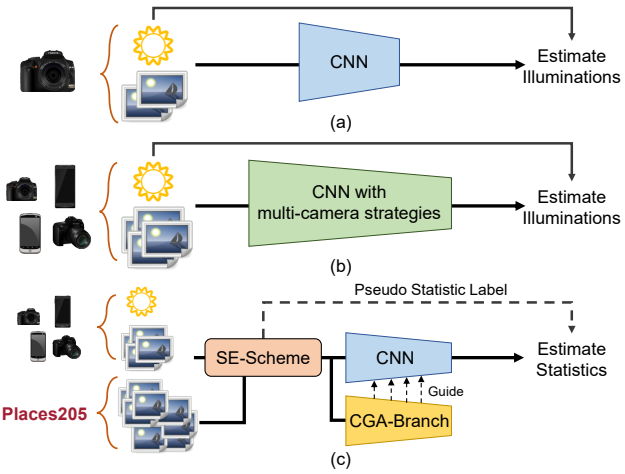

Figure 1: The difference between TLCC and existing methods. (a) Classic CNN-based methods are limited to single camera data. (b) Recent methods adopt the multi-camera strategies to improve the network by joint training of multi-camera data. (c) Our TLCC method can leverage both multi-camera RAW data and massive unlabeled sRGB data.

Introduction

Computational Color Constancy (CCC) aims to remove illumination color casts in RAW images, which helps improve accuracy for many downstream tasks, such as visual recognition (Chen et al. 2015), image segmentation (Afifi and Brown 2019b), etc (Diamond et al. 2017; Andreopoulos and Tsotsos 2011). In the past, most statistic-based color constancy approaches utilize image statistics or physical attributes to estimate the illumination color of the scene (Land 1977; Van De Weijer, Gevers, and Gijsenij 2007). However, the statistics for the reflectance distribution are oversimplified, and these methods are arduous to cope with various scenes in the complicated world. Over the years, learning-based color constancy methods have achieved a remarkable improvement (Hu, Wang, and Lin 2017; Barron and Tsai 2017). They can effectively use complex nonlinear functions to extract illumination cues from the single-camera data automatically, thereby fitting rough illumination curves and obtaining strong generalization capabilities to some extent (Fig. 1(a)). However, due to the sensor domain gap (Hernandez-Juarez et al. 2020) and costly data collection (Shi 2010), learning-based methods are limited to single-camera and face seriously insufficient data problem(Xiao, Gu, and Zhang 2020).

To deal with this problem, some methods have been proposed. A common technic is to reduce the model parameters to prevent over-fitting the few images but limit the model capacity (Hu, Wang, and Lin 2017). Recently, to overcome the influence of the sensor domain gap to benefit from multi-camera joint training, several approaches (McDonagh et al. 2018; Xiao, Gu, and Zhang 2020) apply the ideas of multi-domain learning and few-shot learning strategies to extend the training set (Fig. 1(b)). However, these methods are still restricted in RAW datasets with few images and cannot really solve the lack of data problem in the CCC task.

In this work, we present a transfer learning color con-

*Corresponding Author.

stancy (TLCC) method that introduces cross-camera RAW data and massive unlabeled sRGB data to support the model training (Fig. 1(c)), thereby alleviate the problem of insufficient data. Firstly, the proposed Statistic Estimation Scheme (SE-Scheme) converts the camera-related illumination labels and the input images into the camera-agnostic statistic form. It avoids fitting the specific sensor and makes the multi-camera labels' distribution overlapped each other, thus benefits multi-camera regression. Moreover, the SE-Scheme can generate pseudo statistic labels for approximate color-balanced sRGB data, so as to truly break data limitation for the CCC task. Secondly, to efficiently transfer the scene information from sRGB to RAW data, we propose the Color-Guided Adaptation Branch (CGA-Branch) that adaptively regularizes the features of the backbone to reduce the huge differences in data. Specifically, the proposed branch extracts the image-specific color features by predicting the Color Space Transform Matrix (CSTM) from the RGB-uv histogram (Afifi and Brown 2019a). Then utilize this meaningful feature as conditions to drive the proposed Color-Guided Instance Normalization (CGIN) module, whose parameters can adaptively scale and shift the feature maps, leading the sRGB knowledge to generalize to RAW data.

In summary:

- The proposed TLCC method included SE-Scheme and CGA-Branch breaks data limitation, obtains large model capacity and rich scene information.
- The proposed SE-Scheme avoids fitting the specific sensor for RAW data and provides pseudo statistic labels for sRGB data, thus alleviating insufficient data problem in the training stage.
- The proposed CGA-Branch extracts an image-specific color feature to regularize the backbone's feature map, which realizes efficient transfer learning for sRGB data.
- The experimental results show that our proposed TLCC method achieves state-of-the-art performance on two popular annotated benchmarks.

Related Work

Overview for Color Constancy The CCC approaches concern with estimating illumination on RAW images, which is usually divided into two categories: the statistics-based methods (Land 1977; Cheng, Prasad, and Brown 2014) and the learning-based methods (Barron 2015; Qiu, Xu, and Ye 2020). The former generally estimate the illumination by building assumptions on statistics of scene information. Despite theirs fast speed and insensitivity to cameras, the simple assumptions can not fit the complex real-world well and thus limit performance. Recently, many learning-based color constancy methods based on the convolutional neural network (CNN) have been proposed. They have a stronger generalization ability, and the main difference lies in the regression strategies used: (I) Predict illumination directly (Hu, Wang, and Lin 2017; Yu et al. 2020); (II) Transform to a 2D spatial localization task (Barron 2015; Barron and Tsai 2017); (III) Learn the features of potential achromatic pixels (Bianco and Cusano 2019; Qiu, Xu, and Ye 2020). (IV) Combine with Contrastive Learning (Xu et al. 2020; Lo et al. 2021). However, due to the sensor domain gap (Hernandez-Juarez et al. 2020) and costly data collection (Shi 2010), these learning-based methods only bring improvement in single-camera settings. So that most of them face the problem of lacking annotation data, resulting in shallow model capacity and degradation in multi-camera settings.

Color constancy with insufficient data To remedy the problem of insufficient data, some approaches have been proposed. The most widely used technics in deep learning are data augmentation and the pre-trained models fine-tuning (Hu, Wang, and Lin 2017). However, the former cannot increase the diversity of scene information and cannot guarantee the model's effectiveness in some scenarios. For the latter, due to the illumination information is distorted in the early pre-trained layers (Laakom et al. 2020), the priors from classification tasks can not be transferred effectively. Recently, several approaches (McDonagh et al. 2018; Xiao, Gu, and Zhang 2020) combined with other areas' ideas have been introduced to alleviate the lack of data. McDonagh et al. (McDonagh et al. 2018) utilized the concept of color temperature and Model-Agnostic Meta-Learning algorithm to obtain a meta-model that can be adapted to a new device with few training samples. MDLCC (Xiao, Gu, and Zhang 2020) regarded different cameras as different domains and set parameters to learn each camera's public and private features, which enabled to jointly train with multi-camera data and overcame the data limit in single-camera. SIIE (Afifi and Brown 2019a) reduced the difference between cameras by mapping input images to the public workspace. However, these methods are still learning limited knowledge from RAW datasets and cannot really solve the insufficient data problem. Considering the public sRGB scene recognition datasets, such as Place205 (Zhou et al. 2014), are 2-4 orders of magnitude larger than the RAW dataset, a bold idea is transferring the rich scene information from sRGB to RAW data. Motivated by this idea, Bianco et al. (Bianco and Cusano 2019) designed a network to detect achromatic pixels in gray-scale images, enabling to pre-train the model on the sRGB datasets and then presented the approximate loss to finetune on the RAW datasets. But the gray-scale image discards the color information, making the achromatic pixels detection become another challenging ill-posed problem.

This paper presents a more effective model through transfer learning and scenario statistics information. Unlike the camera-related illumination labels, our method produces the camera-agnostic statistic labels, which allow multi-camera RAW data and sRGB data to train jointly. Besides, the proposed CGA-branch processes the color feature, rather than discards it, to achieve efficacious knowledge transfer.

Transfer Learning in Color Constancy

We present a TLCC method that performs two stages to alleviate the insufficient data problem: (i) The SE-Scheme converts the illumination labels into statistic form and provides the pseudo label for sRGB data to accomplish data extension. (ii) The CGA-Branch extracts meaningful color features to reduce the data difference, which helps to apply massive sRGB data to the CCC task effectively.

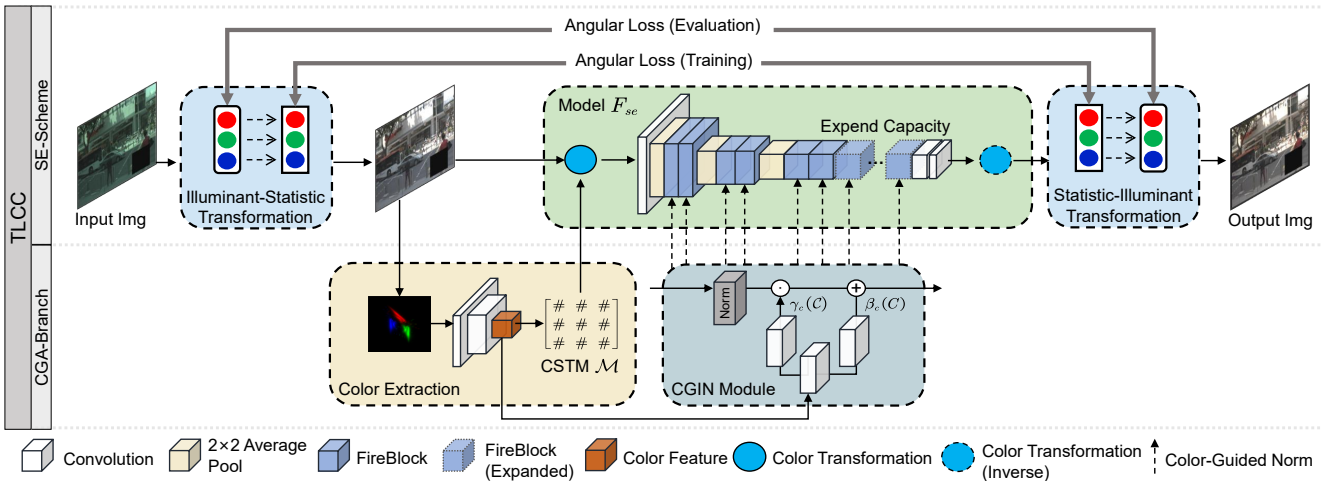

Figure 2: Illustration of the TLCC. SE-Scheme maps the images and illumination labels into statistic form during training and transforms them back into illumination form in evaluation. The proposed CGA-Branch extracts the image-specific color features by predicting the CSTM from the RGB-uv histogram. And drive this color feature to adaptively regularize the feature map by CGIN module. The main body of network F_{se} is based on FC4.

Preliminary of Image Formulation

Assume the RAW image is captured by a digital trichromatic camera, and the image composition can be described by the Lambert model (Barnard 1999). Under the assumption of the uniform illumination source and Von Kries coefficient law (Brainard and Wandell 1986), the RAW image can be modeled as:

$$I(x) = R(x) \circ L \circ C \tag{1}$$

where $I(x) \in \mathbb{R}^3$ is the raw intensity vector at the pixel x, $R(x) \in \mathbb{R}^3$ is the RGB value of reflectance under canonical illumination, $L \in \mathbb{R}^3$ is RGB vector of the arbitrary global illumination, $C \in \mathbb{R}^3$ represents the spectral sensitivity of camera sensor and $\circ$ is Hadamard product.

Different from the RAW image located in the camera-specific color space, the sRGB data is rendered in the standard RGB space by camera pipeline, which is 8 bits and the most applied image data type in real life.

Statistic Estimation Scheme

, we present a new color constancy scheme from the statistic perspective. To derive SE-Scheme, we start from Grey Edge (GE) algorithm (Van De Weijer, Gevers, and Gijsenij 2007) that is the most general statistic-based method. Based on Eq. 1, the common GE framework can be written as:

$$\begin{aligned} h(I,n,\sigma,p) &= \left(\int \left| \frac{\partial^n I^\sigma(x)}{\partial x^n} \right|^p dx \right)^{1/p} \\ &= \left(\int \left| \frac{\partial^n R^\sigma(x)}{\partial x^n} \right|^p dx \right)^{1/p} \circ L \circ C \\ &= h(R,n,\sigma,p) \circ L \circ C \end{aligned} \tag{2}$$

where $h(I,n,\sigma,p)$ denotes the illumination estimated by GE for image I, n is color derivative order, σ is the scale of gaussion filter, p denotes Minkowski norm. The GE estimation includes statistics on surface reflectance, illumination L and sensors C, where the L and C can be eliminated by the Von Kries model (Brainard and Wandell 1986) as:

$$\begin{aligned} & I(x) \circ h(I,n,\sigma,p)^{-1} \\ &= R(x) \circ (L \circ C) \circ (L \circ C)^{-1} \circ h(R,n,\sigma,p)^{-1} \\ &= R(x) \circ h(R,n,\sigma,p)^{-1} \end{aligned} \tag{3}$$

where $(\cdot)^{-1}$ is taking element-wise reciprocal. According to Eq. 3, we get the camera-agnostic result that only contains surface reflectance R. However, due to over-simplified statistics for the reflectance distribution, the newly added item $h(R,n,\sigma,p)^{-1}$ greatly disturbs the accuracy of GE.

In our scheme, we combine the statistic perspective with the CNN, directly using the strong nonlinear capabilities to fit the distribution of the item $h(R,n,\sigma,p)^{-1}$, so as to eliminate its adverse effect. The CNN F_{se} can be modeled as:

$$\hat{h(R,n,\sigma,p)^{-1}} = F_{se}(R \otimes h(R,n,\sigma,p)^{-1};\theta) \tag{4}$$

where $\otimes$ denotes the Hadamard product shared by all pixels, θ is the parameters of the network, $h(R,n,\sigma,p)^{-1}$ called statistic labels. For comparability, the F_{se} is based on FC4 (Hu, Wang, and Lin 2017).

To support this model, the proposed SE-Scheme builds a new training and validation process. Specifically, according to Eq. 2 and Eq. 3, we transform the original RAW images and illumination labels into the statistic form by multiplying $h(I,n,\sigma,p)^{-1}$, respectively. In training, we feed the transformed images into the network F_{se} to regress the statistic labels and backward the losses. During validation, we transform the estimations back to the illumination form to maintain consistency with the CCC task by removing the previously multiplied term $h(I,n,\sigma,p)^{-1}$.

Unlike the common learning-based scheme (Hu, Wang, and Lin 2017; Barron and Tsai 2017), our SE-Scheme inherits the advantages of statistic-based methods that make the whole training process irrelevant to sensors, thereby our

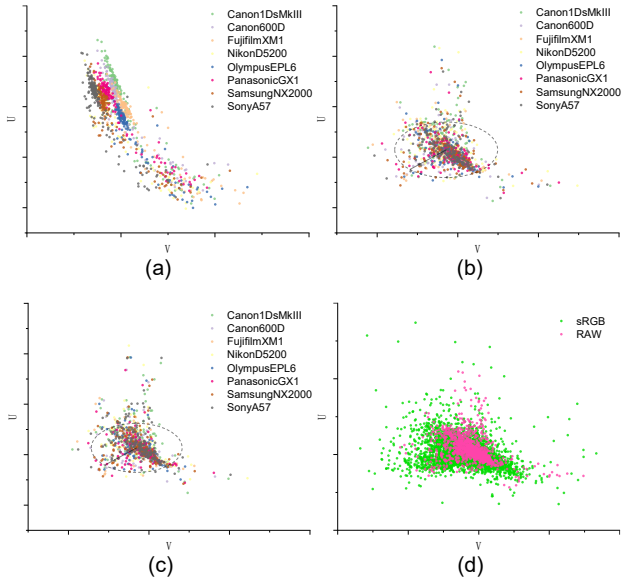

Figure 3: Comparison of the label distributions in UV chromaticity space (Barron 2015) on the eight different cameras. The x- and y-axis represent coordinate V and U, respectively. (a) The original illumination annotations. (b) The statistic labels. The dashed circle with a radius of 0.311 includes the 90% points. (c) The statistic labels are mapped in the temporary color space, and the radius reduces to 0.295. (d) Comparison of all RAW data and sRGB data.

model can adapt to the mixed training of multi-camera RAW data. We visualize the comparison between illumination labels and statistic labels: Since each sensor has a difference in sampling Planck locus, the distribution of multi-camera labels looks like a rainbow (Fig. 3(a)). In contrast, our statistic labels treat each channel of images as a set of sequences and calculate from a statistic perspective. Thus similar scenes produce similar labels, leading to aggregated distribution without any obvious domain gap (Fig. 3(b)), which demonstrates our statistic labels are camera-agnostic. When the label distributions of different cameras tend to be concentrated, the model will benefit from the multi-camera settings.

In addition, our SE-Scheme also has the ability to extend to unlabeled sRGB data. Concretely, we assume the approximate color-balanced sRGB images only contain surface reflectance R so that they can generate the pseudo statistic labels $h(R, n, \sigma, p)^{-1}$ for themselves. We show the distribution of sRGB almost covers the RAW's, which indicates there is no gap in regression targets between them (Fig. 3(d)). We set their initial label as $\mathbf{1}$, so that they can share the same preprocess steps as RAW data does to participate in the training, thereby promoting the model to obtain rich scene information.

Color-Guided Adaption Branch

Although we align the distribution of labels between sRGB and RAW data, the CCC task is also sensitive to color information (Barron 2015; Afifi and Brown 2019a), resulting in inconsistent image feature distribution. Hence, we present the CGA-Branch that extracts meaningful color features by predicting the CSTM and then drives CGIN to adaptively reduce the data difference in each layer.

Color Extraction To extract the meaningful and effective color feature, different from simply obtaining color embedding through CNN (Barron 2015), we set up an additional branch to estimate a CSTM (Afifi and Brown 2019a) that maps the input image into a temporary color space. In order to reduce the error, the learning of CSTM will prompt all images to map from the private color space into the public color space, hence making the feature image-specific, which provides unique information for adaption. Fig. 3(c) shows a more aggregated distribution of labels is provided in the temporary color space: the radius of the 90% points is reduced from 0.311 to 0.295, which indicates that the CSTM is also beneficial to regression. Besides, we convert the input image to parameterized RGB-uv histogram (Afifi and Brown 2019a) that only reserves 2-dimensional color information to ensure the feature is just color-related. This branch can be modeled as:

$$\mathcal{D} = F_{extract}(F_{rgb2uv}(R \otimes h(R, n, \sigma, p)^{-1})) \tag{5}$$

$$\mathcal{M} = \frac{F_{matrix}(\mathcal{D})}{Z} \tag{6}$$

where $F_{rgb2uv}(\cdot)$ is the conversion to RGB-uv histogram, $F_{extract}(\cdot)$ and $F_{matrix}(\cdot)$ contain 3 convolution layers and a single fully connected layer respectively, $\mathcal{M} \in \mathbb{R}^{3\times3}$ denotes the CSTM, Z is a factor that normalizes the matrix to make each row has unit Manhattan norm and $\mathcal{D}$ is our sought meaningful color feature.

During training, we input the mapped image into the network, and the corresponding prediction must map back be to the original color space to maintain consistency with the labels. So the estimation from Eq. 4 becomes $\hat{H}$:

$$\hat{H} = \mathcal{M}^{-1} \cdot F_{se}(\mathcal{M} \times (R \otimes h(R, n, \sigma, p)^{-1})) \tag{7}$$

Algorithm 1: Statistic Estimation Scheme

Input: the RAW dataset $\mathbf{D}_{raw} = \{(x_i, y_i)\}_{i=1}^{m}$ and the approximate color-balanced sRGB dataset $\mathbf{D}_{srgb} = \{(x_i, \mathbf{1})\}_{i=1}^{k}$, where $k \gg m$, model F_{se} with parameters θ, learning rate η.
Output: Trained model parameters.

Initialize model parameters θ.
Divide $\mathbf{D}_{raw}$ into training set $\mathbf{T}_{raw}$ and validation set $\mathbf{V}_{raw}$.
for each pair $(x_i, y_i) \in \{\mathbf{D}_{srgb}, \mathbf{T}_{raw}\}$ **do**
  $s_i \leftarrow h(x_i, n, \sigma, p)^{-1}$
  $input_i \leftarrow x_i \otimes s_i$
  $label_i \leftarrow y_i \circ s_i$
  $pred_i \leftarrow F_{se}(input_i; \theta)$
  $\mathcal{L}_i \leftarrow Loss(pred_i, label_i)$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_i$
end for
for each pair $(x_i, y_i) \in \mathbf{V}_{raw}$ **do**
  $s_i \leftarrow h(x_i, n, \sigma, p)^{-1}$
  $infer_i \leftarrow F_{se}(x_i \otimes s_i; \theta)$
  $\mathcal{L}_i \leftarrow Loss((infer_i \circ s_i^{-1}), y_i)$
end for
return θ

Methods	ColorChecker (Shi 2010)					NUS 8 (Cheng, Prasad, and Brown 2014)				
	Mean	Med.	Tri.	Best 25%	Worst 25%	Mean	Med.	Tri.	Best 25%	Worst 25%
Statistics-based										
White-Patch (Land 1977)	7.55	5.68	6.35	1.45	16.12	9.91	7.44	8.78	1.44	21.27
Grey-World (Buchsbaum 1980)	6.36	6.28	6.28	2.33	10.58	4.14	3.20	3.39	0.90	9.00
Shades-of-Gray (Finlayson and Trezzi 2004)	4.93	4.01	4.23	1.14	10.20	3.67	2.94	3.03	0.98	7.75
1st-order Gray-Edge (Van De Weijer, Gevers, and Gijsenij 2007)	5.33	4.52	4.73	1.86	10.03	3.35	2.58	2.76	0.79	7.18
Bayesian (Gehler et al. 2008)	4.82	3.46	3.88	1.26	10.49	3.50	2.36	2.57	0.78	8.02
Natural Image Statistics (Gijsenij and Gevers 2010)	4.19	3.13	3.45	1.00	9.22	3.45	2.88	2.95	0.83	7.18
LSRS (Gao et al. 2014)	3.31	2.80	2.87	1.14	6.39	3.45	2.51	2.70	0.98	7.32
Grey Pixel (Yang, Gao, and Li 2015)	4.60	3.10	-	-	-	3.15	2.20	-	-	-
GI (Qian et al. 2019)	3.07	1.87	2.16	0.43	7.62	2.91	1.97	2.13	0.56	6.67
Learning-based										
Regression Tree (Cheng et al. 2015)	2.42	1.65	1.75	0.38	5.87	2.36	1.59	1.74	0.49	5.54
DS-Net (Shi, Loy, and Tang 2016)	1.90	1.12	1.33	0.31	4.84	2.24	1.46	1.68	0.48	6.08
FFCC (Barron and Tsai 2017)	1.80	0.95	1.18	**0.27**	4.65	1.99	1.31	1.43	**0.35**	4.75
SqueezeNet-FC4 (Hu, Wang, and Lin 2017)	1.65	1.18	1.27	0.38	3.78	2.23	1.57	1.72	0.47	5.15
Meta-AWB (McDonagh et al. 2018)	2.57	1.84	1.94	0.47	6.11	1.89	1.34	1.44	0.45	4.28
Quisa-U CC (Bianco and Cusano 2019)	2.91	1.98	-	-	-	1.97	1.41	-	-	-
SIIE (Afifi and Brown 2019a)	2.77	1.93	-	0.55	6.53	2.05	1.50	-	0.52	4.48
Multi-Hypothesis-CC (Hernandez-Juarez et al. 2020)	2.10	1.32	1.53	0.36	5.10	2.35	1.55	1.73	0.46	5.62
IGTN (Xu et al. 2020)	1.58	**0.92**	-	0.28	3.70	1.85	**1.24**	-	0.36	4.58
MDLCC (Xiao, Gu, and Zhang 2020)	1.58	0.95	1.11	0.37	3.77	1.78	1.29	1.40	0.42	3.97
TLCC (Our proposed)	**1.51**	0.98	**1.07**	0.33	**3.52**	**1.61**	1.27	**1.33**	0.44	**3.35**

Table 1: Comparison and evaluation with other color constancy methods in CCD and NUS 8 in units of degrees.

Color-Guided Insance Normalization To reduce the difference between sRGB and RAW data, we plug in the proposed CGIN module after each convolution layer. It mainly adopts the idea of adaptive instance normalization (Ulyanov, Vedaldi, and Lempitsky 2016; Huang and Belongie 2017; Kim et al. 2020) that drives the image-specific color features $\mathcal{D}$ to normalize the feature map adaptively and produces affine transform parameters for each channel, thereby reducing the difference in the feature extraction stage for different data. The process can be expressed as:

$$x_c^{norm} = \gamma_c(\mathcal{D})(\frac{x_c - \mu_c}{\sigma_c}) + \beta_c(\mathcal{D}) \tag{8}$$

where c is the number of channels, x_c denotes the input feature, μ_c and σ_c denote the mean and standard deviation of x_c respectively, the rescale $\gamma_c(\mathcal{D})$ and shift $\beta_c(\mathcal{D})$ parameters are guided by the meaningful color features $\mathcal{D}$ through two simple convolution layers. Due to the backbone treats different types of data in the same way, it extracts inconsistent image feature distribution. Hence we utilize the image-specific color features as conditions to compensate for the huge color gap caused by the sRGB data, thereby guiding the reduction of the data difference.

Experiments

Implementing Details

Angular loss In the CCC task, the angular loss is commonly adopted as an evaluation criterion between prediction $\hat{p}$ and ground truth p (Hordley and Finlayson 2004; Barron and Tsai 2017; Qian et al. 2019):

$$AngularLoss(\hat{p}, p) = \frac{180}{\pi} arccos(\hat{p} \cdot p) \tag{9}$$

During training, our proposed statistic estimation scheme directly uses $\hat{H}$ to calculate the angular loss. In the test phase, we map the statistic estimation back into illumination form to calculate the loss through Eq. 5 to maintain consistency with the illumination estimation task.

Training Detail The hyper-parameters n, σ, p of the GE are set as $0, 1, 0$, respectively. And for the function F_{rgb2uv}, the image is resized into 150×150. We employ the Adam (Kingma and Ba 2014) solver as the optimizer and set the learning rate to 1×10^{-4}. We train the model for 1,500 epochs with image size 512×512 and batch size 16. For the first 250 epochs, the sRGB data is firstly fed into the model, followed by RAW data. For the rest epochs, we only use the RAW dataset to fine-tuning the model.

Datasets and Settings

We verify the effectiveness of our proposed method on two public color constancy datasets:

- The reprocessed Color Checker dataset (CCD) (Gehler et al. 2008; Shi 2010) includes indoor and outdoor scenes taken by two cameras, comprising 568 images in total.
- The NUS 8-Camera dataset (NUS 8) (Cheng, Prasad, and Brown 2014) is the multi-camera dataset, consisting of 1736 images taken by 8 different cameras in 260 scenes.

For each RAW dataset, the calibration objects have been masked out, followed by black-level subtraction, saturation pixel clip, and gamma correction. We adopt three-fold cross-validation for each RAW dataset on all experiments followed by the previous works in (Barron 2015; Hernandez-Juarez et al. 2020). For the training, the CCD and the NUS 8 dataset are mixed together as the RAW training set. We further report five standard metrics (Hu, Wang, and Lin 2017; Hernandez-Juarez et al. 2020): mean, median, tri-mean of all angular errors, the mean of the best 25% of angular errors, and the mean of the worst 25% of angular errors.

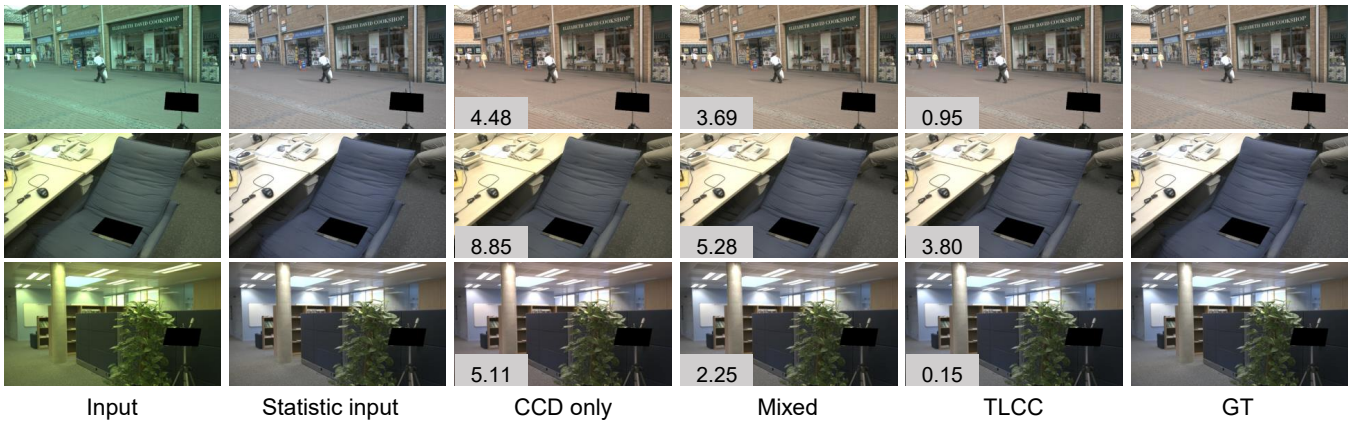

Figure 4: Visualization of proposed TLCC results by CCD only, CCD and NUS 8 mixed, and mixed dataset with additional 12,000 sRGB data. Images are performed gamma correction for visualization.

For the sRGB dataset, We mainly use Place205 (Zhou et al. 2014), which includes 205 scene categories for a total of 2.5 million images. This dataset mainly serves as a benchmark for scene recognition systems. In our experiment, we selected 12,000 approximately color-balanced images from Place205 as the sRGB training set.

Comparison with State-of-the-art Methods

In this section, we compare the quantitative results of our proposed method with other color constancy algorithms over the two popular benchmarks as CCD and NUS 8. The results are listed in Table.1. Our method shows strong competitiveness across five metrics, especially far ahead on the mean and worst 25% metrics, with improvements of 4.4% and 4.9% in CCD and 9.6% and 15.6% in NUS 8 respectively. It indicates that our TLCC method can greatly improve the accuracy of difficult test samples and increase the lower limit of the model. The tri-mean metric is slightly better than all the other methods, and most of the remaining metrics are also in the top three, which shows that we are comparable to other state-of-the-art methods on simple samples. The table also shows a similar rank on both benchmarks means that our method has strong stability. Furthermore, it is worth mentioning that our proposed method is superior to our baseline (FC4) in all metrics, which further proves the effectiveness of our proposed method.

Ablation Study and Analysis

In this section, we perform ablation experiments to evaluate the effectiveness of TLCC's architecture. We conduct nine groups of comparative experiments on the CCD and use the average pooling version of FC4 (the main body of F_{se}) as the baseline. The results are shown in Table 2.

The experiments (1) to (3) show the comparisons when using the proposed SE-Scheme alone. We mainly experiment with two cases of using RAW data only and adding extra sRGB data. Compared with the baseline that adopts the illumination annotations, the SE-Scheme provides the statistic labels with a more aggregated distribution. We see that experiment (2) performs slightly worse than the baseline when using RAW data only. While adding a large number of sRGB images, the performance turns to exceed the baseline in the experiment (3), indicating SE-Scheme with sRGB images is indeed beneficial to the CCC task. These experiments show that it not only obtains better performance under multi-camera settings but also achieves our goal – introducing sRGB images into RAW data to improve performance.

Ablation Study	Mean	Med.	Tri.	Best 25%	Worst 25%
(1) FC4	1.81	1.32	1.42	0.42	4.13
(2) FC4 w SE-Scheme (G)	1.84	1.36	1.50	0.39	4.08
(3) FC4 w SE-Scheme	1.73	1.47	1.45	0.41	3.63
(4) FC4 w IN	1.83	1.31	1.40	0.46	4.09
(5) TLCC w/o CGIN	1.65	1.14	1.22	0.49	3.60
(6) TLCC w/o CSTM	1.58	1.05	1.14	0.37	3.65
(7) TLCC w/o SE-Scheme	1.81	1.42	1.47	0.51	3.83
(8) TLCC (Full)	**1.51**	**0.98**	**1.07**	**0.33**	**3.52**

Table 2: The ablation study of our proposed structure on CCD. The G denotes the control group without sRGB data.

The experiments (4) to (8) show the ablation study of the CGA-Branch's structure. We test the effect of each module separately and serve the FC4 with Instance Normalization (IN) (Ulyanov, Vedaldi, and Lempitsky 2016) as a comparison of the CGIN module. When simply adding the IN module, it does not consider that the different data domain affects the overall offset of the feature, which slightly reduces the accuracy. The proposed CGIN module solves this problem by extracting image-specific color features to generate the scale and shift parameters adaptively. The experiment (6) shows that the CSTM plays a crucial role in further promoting joint training of multi-type data. Without the CSTM, the model will not align in image level and work in chaotic color spaces, which increases the burden of feature level alignment. Moreover, the experiment (7) shows that CGA-Branch working alone will face the sensor gap and unavailable sRGB data, causing the performance degradation to the baseline. Finally, compared with experiment (3), the full structure of the TLCC shows the proposed CGA-Branch further realizes efficient multi-camera learning and

Training Data scope	#	Mean	Med.	Tri.	Best 25%	Worst 25%
FC4						
CCD only	0	1.81	1.32	1.42	0.42	4.13
Mixed	0	1.86	1.41	1.49	0.54	3.98
CCD only	1	1.98	1.51	1.62	0.54	4.24
Mixed	1	2.00	1.50	1.62	0.58	4.11
TLCC						
CCD only	0	1.74	1.28	1.42	0.47	3.81
Mixed	0	1.65	1.21	1.28	0.41	3.57
Mixed + 3,000 sRGB	0	1.58	1.10	1.25	0.39	3.55
Mixed + 6,000 sRGB	1	1.55	1.05	1.13	0.37	3.55
Mixed + 9,000 sRGB	1	1.53	1.07	1.13	0.33	3.55
Mixed + 12,000 sRGB	2	1.51	0.98	1.07	0.33	3.52

Table 3: Comparison of data scope and model capacity on CCD. # represents the number of additional basic blocks.

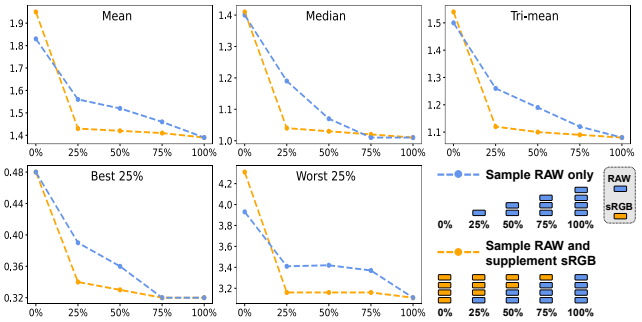

Figure 5: Evaluation for the effect of transfer learning on Cube+ dataset. The ordinate and abscissa are angular error and sampling ratio of RAW data, respectively.

sRGB knowledge transfer.

Discussion on Data Extension and Model Capacity

As aforementioned, our method benefits from massive unlabeled sRGB data and multi-camera RAW data, so we perform two sets of experiments to explore the impact of data expansion on model capacity and performance. Specifically, the first set of experiments display the effect of deeper model depth and more cross-camera data under the FC4 environment. We increase the model capacity by stacking more basic blocks in the backbone, and the specific location is marked in Fig. 2. And we compare two cases of data scope: CCD only, CCD and NUS 8 mixed. The second set of experiments are focused on our proposed method. We add 3,000, 6,000, 9,000, 12,000 sRGB images based on the mixed RAW dataset respectively. To prevent the number of images from reaching the upper limit of the model, we deepen the model at most 2 layers. The results are listed in Table 3.

The first set of experiments show that when the model is built deeper, or the data becomes mixed, the performance of FC4 declines, which demonstrates insufficient data and sensor domain gap cause the model degradation. In contrast, the experiments on TLCC show that the performance is greatly improved when mixing data from different cameras for training, which indicates our method can overcome the sensor domain gap. Meanwhile, with the increasing of sRGB data, TLCC is still capable of boosting when adding more basic blocks, proving that our method can also break the model capacity limitation. Thereby the CCC task can use more classic networks without worrying about overfitting. We further provide some qualitative results in Fig. 4.

Effectiveness of Transfer Learning

We further discuss the effectiveness of transfer learning from sRGB to RAW data in this section. The main experimental dataset is Cube+ (Banić, Koščević, and Lončarić 2017) that contains the 1707 RAW images and the corresponding white-balanced sRGB images. A third of RAW images are served as the test set, and the rest are the training set. We implement two variants: (1) Sampling RAW training set only; (2) Sampling RAW and Supplementing sRGB: Replacing the remaining unsampled RAW data with sRGB data and supplement into the sampled training set. Concretely, sampling can reduce the training set and affect the effect of the model. We supplement the same scenes sRGB data to prove the effectiveness of transfer learning from the sRGB data to RAW data. We set up five experiments: variant1 samples 0%, 25%, 50%, 75%, 100% RAW training set, and variant2 supplements remaining 100%, 75%, 50%, 25%, 0% sRGB data. The performance of the test set is shown in Fig. 5.

As can be observed, compared with the performance of variant1 gradually improved as RAW images increased, the variant2 directly reaches the almost best performance after acquiring a small number of RAW images in the target domain. While the RAW image increasing from 25% to 100%, the performance of variant2 is only slightly improved, which explains that we can learn from a large amount of sRGB scene information and then transfer it to the target RAW domain that suffers insufficient data immediately.

Extension

In this paper, our proposed SE-Scheme is based on the Von Kries model, and it still has the potential to adapt to more complex image correction models, such as the Diagonal-offset model (Shafer 1985). The extra offset term can be eliminated by calculating the image's first derivative, and the rest steps are the same as this paper does. In general, as long as we remove the factors that unrelate to surface reflectance R, the SE-Scheme can be built.

Conclusion

This paper presents the TLCC method to alleviate the CCC task's insufficient data problem by introducing the multi-camera RAW and sRGB data. We achieve this by regressing the proposed statistic label and driving image-specific color features to reduce data difference adaptively. The experimental results show that the proposed method favors a deeper model with multi-camera settings and achieves state-of-the-art performance on two public datasets. We also evaluate the effectiveness of transfer learning, which shows that we can leverage massive sRGB scene information to transfer into the small RAW datasets. In future work, we plan to extend our method to the more complex reflectance model and image correction model, which are closer to the real world.

Acknowledgments

This work was supported by the national key R & D program intergovernmental international science and technology innovation cooperation project (2021YFE0101600).

References

Afifi, M.; and Brown, M. S. 2019a. Sensor-independent illumination estimation for DNN models. *arXiv preprint arXiv:1912.06888*.

Afifi, M.; and Brown, M. S. 2019b. 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*, 243–252.

Andreopoulos, A.; and Tsotsos, J. K. 2011. On sensor bias in experimental methods for comparing interest-point, saliency, and recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1): 110–126.

Banić, N.; Koščević, K.; and Lončarić, S. 2017. Unsupervised learning for color constancy. *arXiv preprint arXiv:1712.00436*.

Barnard, K. 1999. *Practical colour constancy*. Simon Fraser University Burnaby.

Barron, J. T. 2015. Convolutional color constancy. In *Proceedings of the IEEE International Conference on Computer Vision*, 379–387.

Barron, J. T.; and Tsai, Y.-T. 2017. Fast fourier color constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 886–894.

Bianco, S.; and Cusano, C. 2019. Quasi-unsupervised color constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12212–12221.

Brainard, D. H.; and Wandell, B. A. 1986. Analysis of the retinex theory of color vision. *JOSA A*, 3(10): 1651–1661.

Buchsbaum, G. 1980. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1): 1–26.

Chen, Y.-H.; Chao, T.-H.; Bai, S.-Y.; Lin, Y.-L.; Chen, W.-C.; and Hsu, W. H. 2015. Filter-invariant image classification on social media photos. In *Proceedings of the 23rd ACM international conference on Multimedia*, 855–858.

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. *JOSA A*, 31(5): 1049–1058.

Cheng, D.; Price, B.; Cohen, S.; and Brown, M. S. 2015. Effective learning-based illuminant estimation using simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1000–1008.

Diamond, S.; Sitzmann, V.; Boyd, S.; Wetzstein, G.; and Heide, F. 2017. Dirty pixels: Optimizing image classification architectures for raw sensor data. *arXiv e-prints*, arXiv–1701.

Finlayson, G. D.; and Trezzi, E. 2004. Shades of gray and colour constancy. In *Color and Imaging Conference*, volume 2004, 37–41. Society for Imaging Science and Technology.

Gao, S.; Han, W.; Yang, K.; Li, C.; and Li, Y. 2014. Efficient color constancy with local surface reflectance statistics. In *European Conference on Computer Vision*, 158–173. Springer.

Gehler, P. V.; Rother, C.; Blake, A.; Minka, T.; and Sharp, T. 2008. Bayesian color constancy revisited. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Gijsenij, A.; and Gevers, T. 2010. Color constancy using natural image statistics and scene semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4): 687–698.

Hernandez-Juarez, D.; Parisot, S.; Busam, B.; Leonardis, A.; Slabaugh, G.; and McDonagh, S. 2020. A multi-hypothesis approach to color constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2270–2280.

Hordley, S. D.; and Finlayson, G. D. 2004. Re-evaluating colour constancy algorithms. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, 76–79. IEEE.

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

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Kim, Y.; Soh, J. W.; Park, G. Y.; and Cho, N. I. 2020. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3482–3492.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laakom, F.; Passalis, N.; Raitoharju, J.; Nikkanen, J.; Tefas, A.; Iosifidis, A.; and Gabbouj, M. 2020. Bag of color features for color constancy. *IEEE Transactions on Image Processing*, 29: 7722–7734.

Land, E. H. 1977. The retinex theory of color vision. *Scientific american*, 237(6): 108–129.

Lo, Y.-C.; Chang, C.-C.; Chiu, H.-C.; Huang, Y.-H.; Chen, C.-P.; Chang, Y.-L.; and Jou, K. 2021. CLCC: Contrastive Learning for Color Constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8053–8063.

McDonagh, S.; Parisot, S.; Zhou, F.; Zhang, X.; Leonardis, A.; Li, Z.; and Slabaugh, G. 2018. Formulating Camera-Adaptive Color Constancy as a Few-shot Meta-Learning Problem. *arXiv preprint arXiv:1811.11788*.

Qian, Y.; Kamarainen, J.-K.; Nikkanen, J.; and Matas, J. 2019. On finding gray pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8062–8070.

Qiu, J.; Xu, H.; and Ye, Z. 2020. Color Constancy by Reweighting Image Feature Maps. *IEEE Transactions on Image Processing*, 29: 5711–5721.

Shafer, S. A. 1985. Using color to separate reflection components. *Color Research & Application*, 10(4): 210–218.

Shi, L. 2010. Re-processed version of the gehler color constancy dataset of 568 images. http://www.cs.sfu.ca/~colour/data/. Accessed: 2010-09.

Shi, W.; Loy, C. C.; and Tang, X. 2016. Deep specialized network for illuminant estimation. In *European conference on computer vision*, 371–387. Springer.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Van De Weijer, J.; Gevers, T.; and Gijsenij, A. 2007. Edge-based color constancy. *IEEE Transactions on image processing*, 16(9): 2207–2214.

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*, 3258–3267.

Xu, B.; Liu, J.; Hou, X.; Liu, B.; and Qiu, G. 2020. End-to-end illuminant estimation based on deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3616–3625.

Yang, K.-F.; Gao, S.-B.; and Li, Y.-J. 2015. Efficient illuminant estimation for color constancy using grey pixels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2254–2263.

Yu, H.; Chen, K.; Wang, K.; Qian, Y.; Zhang, Z.; and Jia, K. 2020. Cascading convolutional color constancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12725–12732.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.