Cascading Convolutional Color Constancy

Huanglin Yu,[1] **Ke Chen,**[1,*] **Kaiqi Wang,**[1] **Yanlin Qian,**[2] **Zhaoxiang Zhang,**[3] **Kui Jia**[1]
[1]South China University of Technology, [2]Tampere University, [3]Chinese Academy of Sciences
{eeyu.huanglin, mswkq}@mail.scut.edu.cn, {chenk, kuijia}@scut.edu.cn
yanlin.qian@tuni.fi, zhaoxiang.zhang@ia.ac.cn

Abstract

Regressing the illumination of a scene from the representations of object appearances is popularly adopted in computational color constancy. However, it's still challenging due to intrinsic appearance and label ambiguities caused by unknown illuminants, diverse reflection properties of materials and extrinsic imaging factors (such as different camera sensors). In this paper, we introduce a novel algorithm – *Cascading Convolutional Color Constancy* (in short, C^4) to improve robustness of regression learning and achieve stable generalization capability across datasets (different cameras and scenes) in a unique framework. The proposed C^4 method ensembles a series of dependent illumination hypotheses from each cascade stage via introducing a weighted multiply-accumulate loss function, which can inherently capture different modes of illuminations and explicitly enforce coarse-to-fine network optimization. Experimental results on the public Color Checker and NUS 8-Camera benchmarks demonstrate superior performance of the proposed algorithm in comparison with the state-of-the-art methods, especially for more difficult scenes.

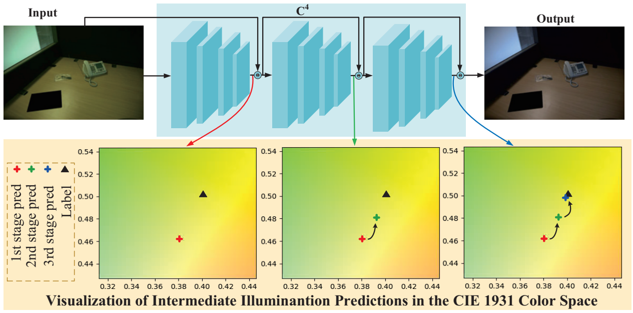


Figure 1: Visualization of the proposed C^4 method in a cascaded structure in the top row, while plots in the bottom row show dependent illumination hypotheses of different cascade stages in our C^4 on an example from the Color Checker dataset. Our C^4 can significantly boost illumination estimation performance in a coarse-to-fine refining manner. More examples are given in the experiment section.

Introduction

The colors present in images are biased by the illumination in addition to the intrinsic reflection properties of scene objects and extrinsic spectral sensitivity across cameras, but they appear to be relatively constant for human visual perception system. Such a property, referred to as color constancy, makes object appearance under diverse lighting sources independent of the casting illumination, which is desired in a large number of high-level vision problems. The color constancy problem can typically be addressed via estimating the color of the illuminant of the scene firstly, which then recovers the canonical colors of scene objects. A large number of computational color constancy algorithms (Qian et al. 2019; Chen et al. 2019; Cheng et al. 2015; Bianco, Cusano, and Schettini 2017; Shi, Loy, and Tang 2016; Hu, Wang, and Lin 2017) rely on accurate and robust illumination predictions and then employ the simple yet effective von Kries model (von Kries 1902) for image correction. Estimating the illumination of an image can be formulated into learning a regression mapping from the imagery representation to its corresponding illumination label. Searching and identifying the best hypothesis of the illumination is not trivial in view of appearance inconsistency and label ambiguity. In addition to unknown surface reflection, the large appearance variation of the captured scene objects can be caused by the sensor sensitivity and also the illuminant spectrum. Specifically, spectral responses of sensors in cameras for color imaging are not consistent across camera models and brands, *e.g.* in the NUS 8-camera dataset (Cheng, Prasad, and Brown 2014) one scene is captured with eight different cameras, and they have visually varying colors for the identical object surface. Consequently, a typical solution is to train camera-specific estimators, which is less efficient and even impractical due to data-demanding characteristics. Very few algorithms (Qian et al. 2019) focus on the challenging camera-agnostic illumination estimation, achieving robust performance. Therefore, the challenge still exits. Most of the existing algorithms (Bianco, Cusano, and Schettini 2017; Barron 2015; Shi, Loy, and Tang 2016; Barron and Tsai 2017;

[*] Corresponding author.

Hu, Wang, and Lin 2017) have been proposed to deal with appearance inconsistency, while very few concern on the challenge caused by the error-prone assumption in practice, *i.e.* one unique spectral illumination exists in the whole scene of each image. In the procedure of label acquisition for color constancy datasets, a Macbeth ColorChecker chart is usually placed in the image, whose colors are recorded as the ground truth illumination, breaking the guarantee: the recorded "ground truth" represents the real global illumination. As a result, the gap between the label and the true scene illumination over spatial regions makes learning a regression more challenging, especially considering data augmentation via patch-based sampling widely adopted in state-of-the-art deep methods (Bianco, Cusano, and Schettini 2017; Shi, Loy, and Tang 2016; Hu, Wang, and Lin 2017). Robustness against object appearance inconsistency and label ambiguity are desired imagery representation properties to learn from imagery observations and illumination labels. To achieve these, we introduce a multiply-accumulate loss function for cascading convolutional color constancy (*e.g.* FC^4 (Hu, Wang, and Lin 2017) in the experiments) to cope with both challenges simultaneously. In details, a series of dependent illumination hypotheses, reflecting different modes of illuminations, are generated via the proposed cascaded model, which are then combined in an ensemble to enforce explicitly coarse-to-fine refinement on illumination hypotheses as Figure 1 shows. The contributions of this paper are three-fold.

- This paper proposes a generic cascaded structure (*i.e.* the multiply-accumulate cascade) on illumination estimation to 1) ensemble multiple dependent illumination hypotheses and 2) achieve coarse-to-fine refinement, via a novel multiply-accumulate loss, which can be readily plugged into other learning-based illumination estimators.
- The proposed C^4 method increases model flexibility via enriching abstract features in a deeper network structure and also discovers latent correlation in the hypothesis space, which alleviates the suffering from ambiguous training samples.
- Extensive experiments on two popular benchmarks show that our C^4 achieves significantly better performance than the state-of-the-art, especially when coping with more difficult scenes.

Source codes and pre-trained models are available at https://github.com/yhlscut/C4.

Related work

Color constancy has been investigated for decades and numerous conventional algorithms are based on low-level imagery statistics, such as White-Patch (Brainard and Wandell 1986), Gray-World (Buchsbaum 1980), Gray-Edge (Van De Weijer, Gevers, and Gijsenij 2007), Shades-of-Gray (Finlayson and Trezzi 2004), Bright Pixels (Joze et al. 2012), Grey Pixel (Yang, Gao, and Li 2015) and Gray Index (Qian et al. 2019). These algorithms are proposed to determine the neutral white color with algorithm-specific assumptions, which encourage direct application to testing images in a learning-free fashion but can be sensitive in practice in consideration of their dependency on statistical distribution of pixel-wise colors, *e.g.* lack of gray pixels with using grey pixels (Yang, Gao, and Li 2015) and state-of-the-art statistical grey index (Qian et al. 2019).

Learning-based methods are a powerful alternative for generating constant colors under a scene illumination, which can be categorized into two groups – gamut mapping (Barnard 2000; Chakrabarti, Hirakawa, and Zickler 2011) and regression learning (Funt and Xiong 2004; Cheng et al. 2015; Qian et al. 2017; Chen et al. 2019; Cardei and Funt 1999; Schaefer, Hordley, and Finlayson 2005; Bianco, Cusano, and Schettini 2017; Barron 2015; Shi, Loy, and Tang 2016; Barron and Tsai 2017; Hu, Wang, and Lin 2017). The former gamut mapping algorithms including edge-based (Barnard 2000), intersection-based (Chakrabarti, Hirakawa, and Zickler 2011) and pixels-based (Chakrabarti, Hirakawa, and Zickler 2011) assume the size of colors under a given illuminant is limited, but will have a variation on observed colors when a deviation in the color of illuminants. Given sufficient labeled training data, a model can be trained to recognize the canonical illumination by mapping from a gamut of a testing image under an unknown illuminant to the canonical gamut, which can thus generate an estimation of the scene illumination.

The latter regression learning-based algorithms aim to learn a direct regression mapping from the imagery representation to its corresponding illumination vector. These methods focus on either designing robust regressors against large feature variation, based on support vector regression (Funt and Xiong 2004), regression trees (Cheng et al. 2015), an ensemble of shallow regressors (Cardei and Funt 1999; Schaefer, Hordley, and Finlayson 2005) or mining inter-dimensional label correlation as structured-output regression (Qian et al. 2016; Chen et al. 2019). Inspired by the recent success of convolutional neural networks on numerous vision tasks, a number of works introduce the 2D convolutional feature encoding into color constancy. (Bianco, Cusano, and Schettini 2017) is the first attempt of deep color constancy, which copes with data sparsity problem demanded by fitting millions of network parameters via patch-based sampling. Convolutional Color Constancy (CCC) (Barron 2015) and Fast Fourier Color Constancy (FFCC) (Barron and Tsai 2017) formulate the problem into a 2D spatial localization task on a 2D log-chroma space, while the difference of both methods lies in better performance and acceleration of the latter benefiting from extra semantic features and the BVM estimation in the frequency domain. In (Hu, Wang, and Lin 2017), a confidence-pooling layer is introduced to automatically feature encoding and discover the location of essential spatial regions for illumination estimation. Existing deep learning methods mainly focus on designing network structure for robust feature encoding against the challenge of inconsistent appearance, but omits to benefit from combining multiple illumination hypotheses in an ensemble to handle ambiguous samples.

Recent DS-Net (Shi, Loy, and Tang 2016) has two expert branches for first generating two hypotheses and then automatically selecting the better one. Similar to our C^4 method,

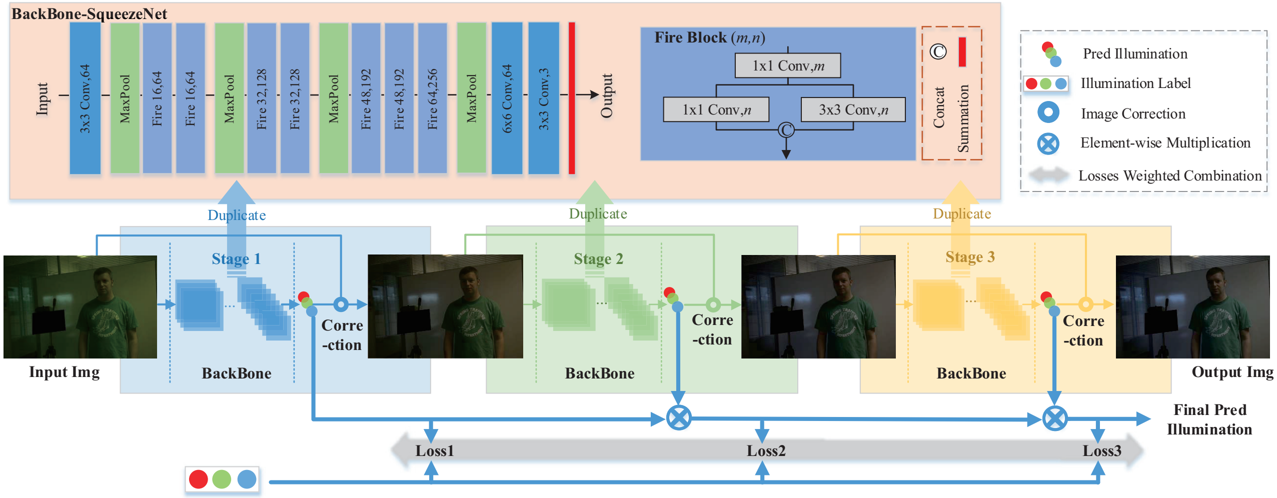


Figure 2: Pipeline of our three-stage C^4 model based on the SqueezeNet backbone.

its motivation is to exploit multiple hypotheses of scene illumination for robust color constancy. However, there are two key differences. Firstly, the DS-Net conducts a discriminative selection instead of jointly learning to discover latent dependency across multiple illumination hypotheses as our C^4 model. Secondly, the DS-Net generates multiple independent illumination hypotheses in parallel, while the proposed C^4 method in a cascading network structure generates dependent hypotheses in serial to explicitly enforce coarse-to-fine refinement. Experiment results in Tables 1 and 2 demonstrate the superiority of our C^4 model to the DS-Net and other state-of-the-art methods.

C^4: Cascading Convolutional Color Constancy

The problem definition of single illumination estimation problem is to predict the illumination vector $\boldsymbol{y} \in \mathbb{R}^3$ from the image $X \in \mathbb{R}^{H\times W\times 3}$. For learning-based illumination estimation, the objective function can be written as the following:

$$\min_{\theta} \ \mathcal{L}(f^{\theta}(X), \boldsymbol{y}), \tag{1}$$

where $f^{\theta}(\cdot) \in \mathbb{R}^3$ is the mapping from the image X to illumination vector $\boldsymbol{y}$, and θ denotes the model parameters of f to be optimized. $\mathcal{L}(\cdot)$ denotes the loss function and the typical loss in illumination estimation is the angular loss (formulated in Equation (3)). During testing, given an input, the trained model $f^{\theta}(\cdot)$ infers the predicted illumination $f^{\theta}(X)$, which is used to generate the color-corrected image. In the context of convolutional color constancy, $f^{\theta}(\cdot)$ is the output of a deep network, while θ denotes the network weights. This section will present an overview of the proposed C^4 algorithm, a novel multiply-accumulate loss, image correction, and implementation details respectively.

Network Structure

The C^4 network consists of multiple stages. Given training pairs $\{X, \boldsymbol{y}\}_i$, $i \in \{1, 2, \cdots, N\}$, in a cascaded structure, $f^{\theta}(\cdot)$ can be decomposed into $f_l(\cdot), l = 1, 2, \ldots, L$, where l and L denote the cascade level and the total number of cascade stages, respectively, with θ omitted for simplicity. We define $f_l f_{l-1}(X)$ as a simpler notation for $f_l(X/f_{l-1}(X))$ (image correction, depicted in Equation (5)). Considering the cascaded structure, now Equation (1) of the three-stage C^4, illustrated in Figure 2, can be written as the following:

$$\min_{\theta} \ \mathcal{L}(f_3 f_2 f_1(X)), \boldsymbol{y}; \theta). \tag{2}$$

In the light of its good performance in illumination estimation, we employ the state-of-the-art CNN model – FC^4 based on the AlexNet and SqueezeNet backbone in (Hu, Wang, and Lin 2017). In details, the FC^4 adopts low-level convolutional layers of off-the-shelf AlexNet and SqueezeNet pre-trained on the ImageNet (Deng et al. 2009), and replaces the remaining layers with two more convolutional layers. Specifically, the AlexNet-FC^4 model keeps all the layers up to the *conv5* layer and replaces the rest fully-connected layers with *conv6* having $6\times6\times64$ convolutional filters and *conv7* ($1\times1\times4$), while the detailed network structure of the SqueezeNet-FC^4 is shown in Figure 2. For both networks, every convolution layers are followed by a ReLU non-linearity, and a dropout with probability 0.5 is added before the last convolutional layer. It is noted that a confidence-weighted pooling layer is followed by the last *conv* layer in original FC^4 to improve robustness against color consistency across spatial regions via suppressing less confident predictions, while our FC^4 model employs a much simpler summation on the output of last *conv* layer to obtain global illumination $\boldsymbol{y}$ (*i.e.*, a red bar in the top row of Figure 2) without hindering the performance.

A Novel Multiply-Accumulate Loss

As mentioned earlier, illumination predictions in different cascade stages are approximate to ground truth illumination, which can be viewed as its different nodes. Different from the DS-Net (Shi, Loy, and Tang 2016) to design a selection mechanism via training another branch to determine the

Table 1: Comparative evaluation on two popular benchmarks. All results reported in this table are in units of degrees.

Methods	NUS 8-Camera (Cheng, Prasad, and Brown 2014)					Color Checker (Shi 2000)				
	Mean	Median	Tri-mean	Best 25%	Worst 25%	Mean	Median	Tri-mean	Best 25%	Worst 25%
Static Methods										
White-Patch (Brainard and Wandell 1986)	10.62	10.58	10.49	1.86	19.45	7.55	5.68	6.35	1.45	16.12
Gray-World (Buchsbaum 1980)	4.14	3.20	3.39	0.90	9.00	6.36	6.28	6.28	2.33	10.58
1st-order Gray-Edge (Van De Weijer, Gevers, and Gijsenij 2007)	3.20	2.22	2.43	0.72	7.69	5.33	4.52	4.73	1.86	10.03
2nd-order Gray-Edge (Van De Weijer, Gevers, and Gijsenij 2007)	3.20	2.26	2.44	0.75	7.27	5.13	4.44	4.62	2.11	9.26
Shades-ofq-Gray (Finlayson and Trezzi 2004)	3.40	2.57	2.73	0.77	7.41	4.93	4.01	4.23	1.14	10.20
General-Gray-World (Barnard, Cardei, and Funt 2002)	3.21	2.38	2.53	0.71	7.10	4.66	3.48	3.81	1.00	10.09
Bright Pixels (Joze et al. 2012)	3.17	2.41	2.55	0.69	7.02	3.98	2.61	-	-	-
Cheng et al.2104 (Cheng, Prasad, and Brown 2014)	2.92	2.04	2.24	0.62	6.61	3.52	2.14	2.47	0.50	8.74
LSRS (Gao et al. 2014)	3.45	2.51	2.70	0.98	7.32	3.31	2.80	2.87	1.14	6.39
Grey Pixel (edge) (Yang, Gao, and Li 2015)	3.15	2.20	-	-	-	4.60	3.10	-	-	-
GI (Qian et al. 2019)	2.91	1.97	2.13	0.56	6.67	3.07	1.87	2.16	0.43	7.62
Learning-based Methods										
Edge-based Gamut (Barnard 2000)	8.43	7.05	7.37	2.41	16.08	6.25	5.04	5.43	1.90	13.58
Bayesian (Gehler et al. 2008)	3.67	2.73	2.91	0.82	8.21	4.82	3.46	3.88	1.26	10.49
MvCA (Chen et al. 2019)	-	-	-	-	-	4.10	2.60	-	-	-
Intersection-based Gamut (Chakrabarti, Hirakawa, and Zickler 2011)	7.20	5.96	6.28	2.20	13.61	4.20	2.39	2.93	0.51	10.70
Pixels-based Gamut (Chakrabarti, Hirakawa, and Zickler 2011)	7.70	6.71	6.90	2.51	14.05	4.20	2.33	2.91	0.50	10.72
Natural Images Statistics (Gijsenij and Gevers 2010)	3.71	2.60	2.84	0.79	8.47	4.19	3.13	3.45	1.00	9.22
Spatio-spectral (GenPrior) (Chakrabarti, Hirakawa, and Zickler 2011)	2.96	2.33	2.47	0.80	6.18	3.59	2.96	3.10	0.95	7.61
Corrected-Moment[1] (19 Color) (Finlayson 2013)	3.05	1.90	2.13	0.65	7.41	2.96	2.15	2.37	0.64	6.69
Corrected-Moment[1] (19 Edge) (Finlayson 2013)	3.03	2.11	2.25	0.68	7.08	3.12	2.38	2.59	0.90	6.46
Exemplar-based (Joze and Drew 2013)	-	-	-	-	-	3.10	2.30	-	-	-
Chakrabarti et al. 2015 (Chakrabarti 2015)	-	-	-	-	-	2.56	1.67	1.89	0.52	6.07
Regression Tree (Cheng et al. 2015)	2.36	1.59	1.74	0.49	5.54	2.42	1.65	1.75	0.38	5.87
CNN (Bianco, Cusano, and Schettini 2017)	-	-	-	-	-	2.36	1.98	-	-	-
CCC (dist+ext) (Barron 2015)	2.38	1.48	1.69	0.45	5.85	1.95	1.22	1.38	0.35	4.76
DS-Net (HypNet+SeNet) (Shi, Loy, and Tang 2016)	2.24	1.46	1.68	0.48	6.08	1.90	1.12	1.33	0.31	4.84
FFCC (Barron and Tsai 2017)	1.99	**1.31**	**1.43**	**0.35**	4.75	1.78	0.96	1.14	0.29	4.62
AlexNet-FC^4 (Hu, Wang, and Lin 2017)	2.12	1.53	1.67	0.48	4.78	1.77	1.11	1.29	0.34	4.29
SqueezeNet-FC^4 (Hu, Wang, and Lin 2017)	2.23	1.57	1.72	0.47	5.15	1.65	1.18	1.27	0.38	3.78
$C^4_{\text{AlexNet-FC4}}$ (ours)	2.07	1.47	1.63	0.48	4.63	1.49	1.03	1.13	0.29	3.52
$C^4_{\text{SqueezeNet-FC4}}$ (ours)	**1.96**	1.42	1.53	0.48	**4.40**	**1.35**	**0.88**	**0.99**	**0.28**	**3.21**

better hypothesis, the proposed cascaded network aims to exploit latent dependency across illumination hypotheses to explicitly enforce coarse-to-fine refinement approaching the ground truth. To this end, we introduce a combined multiply-accumulate loss on all hypotheses to capture their latent correlation to refine illumination hypotheses,which is depicted as the following equation:

$$\mathcal{L} = \sum_{l=1}^{L} \mathcal{L}^{(l)}(\prod_{i=1}^{l} f_i(X_i), \boldsymbol{y}) \tag{3}$$

where $\mathcal{L}^{(l)}$ represents the loss at the l-th cascade stage. Moreover, the proposed loss can alleviate cumulative errors via supervision on intermediate illumination predictions. We also consider its simple weighted extension as

$$\mathcal{L} = \sum_{l=1}^{L} w_l \mathcal{L}^{(l)}(\prod_{i=1}^{l} f_i(X_i), \boldsymbol{y}) \tag{4}$$

where w_l denotes weights for the loss on illumination prediction in the l-th stage and ground truth $\boldsymbol{y}$. We compare the variants of weights in Equation (4) and results are shown in Table 3. The proposed losses are embedded into the deep cascaded network in an end-to-end learning manner as shown in Figure 2.

For large appearance variation and ambiguous labels, a selection or an ensemble of a number of illumination estimators are verified its superior robustness, but it remains challenging to capture latent correlation across illumination hypotheses. The combined loss proposed in this paper is extremely simple yet effective, as the principle of our design can be explained by enforcing each cascaded stage to learn a specific correction pattern to suppress ambiguous hypotheses in previous stages.

Image Correction

With an estimated illumination $\hat{\boldsymbol{y}} = [\hat{y}_r, \hat{y}_g, \hat{y}_b] \in \mathbb{R}^3$ for a biased image $\boldsymbol{X}$ with the trained C^4 model, the canonical colors of scene objects in the image can be recovered under the simplified assumption that each RGB channel can be modified separately (von Kries 1902). In other words, we can obtain the corrected image $\bar{\boldsymbol{X}} \in \mathbb{R}^{H \times W \times 3}$ under the canonical illumination as

$$\bar{\boldsymbol{X}}_j = \boldsymbol{X}_j / y_j \in \mathbb{R}^{H \times W},\ j \in \{R, G, B\}. \tag{5}$$

Implementation Details

In data augmentation, we randomly crop patches from original images with a side length of [0.1, 1] times the shorter side of the original image which are randomly rotated between $-30°$ and $30°$. These patches are then resized into 512 × 512 pixels and finally randomly horizontal flipped with a probability of 0.5. To increase the diversity of limited training data, the illumination labels in each image are scaled by three different random values within the range between 0.6 and 1.4, and pixel-wise scene colors present in the original image are also biased by the randomly generated ratios. We further apply gamma correction to convert linear images into nonlinear images and normalize the values of the images to $[0, 1]$. During training, the ADAM algorithm (Kingma and Ba 2014) is employed to train the model with a fixed batch size (*i.e.* 16 in our experiments), and the learning rate is set to 3×10^{-4} and 1×10^{-4} for our C^4 model based on the SqueezeNet and AlexNet backbone respectively. For computational efficiency and robust performance, we first train the one-stage C^4 for $2,000$ epochs, the learned weights are loaded to each cascade sub-net as initial weights in our three-stage C^4 model for further fine-tuning jointly.

Table 2: Camera-agnostic evaluation. All results are in units of degrees.

Training set	NUS 8-Camera					Color Checker				
Testing set	Color Checker					NUS 8-Camera				
	Mean	Median	Tri-mean	Best 25%	Worst 25%	Mean	Median	Tri-mean	Best 25%	Worst 25%
Static Methods										
White-Path (Brainard and Wandell 1986)	7.55	5.68	6.35	1.45	16.12	9.91	7.44	8.78	1.44	21.27
Gray-World (Buchsbaum 1980)	6.36	6.28	6.28	2.33	10.58	4.59	3.46	3.81	1.16	9.85
1st-order Gray-Edge (Van De Weijer, Gevers, and Gijsenij 2007)	5.33	4.52	4.73	1.86	10.43	3.35	2.58	2.76	0.79	7.18
2nd-order Gray-Edge (Van De Weijer, Gevers, and Gijsenij 2007)	5.13	4.44	4.62	2.11	9.26	3.36	2.70	2.80	0.89	7.14
Shades-of-Gray (Finlayson and Trezzi 2004)	4.93	4.01	4.23	1.14	10.20	3.67	2.94	3.03	0.99	7.75
General-Gray-World (Barnard, Cardei, and Funt 2002)	4.66	3.48	3.81	1.00	10.09	3.20	2.56	2.68	0.85	6.68
Grey Pixel (edge) (Yang, Gao, and Li 2015)	4.60	3.10	-	-	-	3.15	2.20	-	-	-
Cheng et al. 2104 (Cheng, Prasad, and Brown 2014)	3.52	2.14	2.47	0.50	8.74	2.92	2.04	2.24	0.62	6.61
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
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 Methods										
Bayesian (Gehler et al. 2008)	4.75	3.11	3.50	1.04	11.28	3.65	3.08	3.16	1.03	7.33
Chakrabarti et al. 2015 (Chakrabarti 2015)	3.52	2.71	2.80	0.86	7.72	3.89	3.10	3.26	1.17	7.95
FFCC (Barron and Tsai 2017)	3.91	3.15	3.34	1.22	7.94	3.19	2.33	2.52	0.84	7.01
AlexNet-FC^4 (Hu, Wang, and Lin 2017)	3.23	2.57	2.73	0.90	6.70	2.62	2.16	2.25	0.79	5.23
SqueezeNet-FC^4 (Hu, Wang, and Lin 2017)	3.02	2.36	2.50	0.81	6.36	2.40	2.03	2.10	0.70	4.80
$\mathrm{C}^4_{\text{AlexNet-FC4}}$ (ours)	2.85	2.26	2.38	0.76	5.97	2.52	2.07	2.15	0.69	5.20
$\mathrm{C}^4_{\text{SqueezeNet-FC4}}$ (ours)	**2.73**	2.20	2.28	0.72	**5.69**	**2.28**	**1.90**	**1.97**	0.67	**4.60**

Experiments

Datasets and Settings

We conduct experimental evaluation on two public color constancy benchmarks: the NUS 8-Camera dataset (Cheng, Prasad, and Brown 2014) and the re-processed Color Checker dataset (Shi 2000). The NUS 8-camera dataset is composed of 1736 images from 8 commercial cameras, while the Color Checker dataset contains 568 images including indoor and outdoor scenes. All images in both benchmarks are linear images in the RAW format of the acquisition device, each with a Macbeth ColorChecker (MCC) chart, which provides an estimation of illuminant colors.

To prevent the convolutional network from detecting and utilizing MCCs as a visual cue, all images are masked with provided locations of MCC during training and testing. Following (Chen et al. 2019; Qian et al. 2019; Barron 2015), we adopt three-fold cross-validation on both datasets in all experiments.

As suggested in (Hordley and Finlayson 2004) as well as a number of recent works (Chen et al. 2019; Qian et al. 2019; Barron 2015), we use the *angular error* ϵ between the RGB triplet of estimated illuminant $\hat{\boldsymbol{y}}$ and the RGB triplet of the measured ground truth illuminant $\boldsymbol{y}$ as the performance metric denoted as :

$$\epsilon(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \arccos\left(\frac{\hat{\boldsymbol{y}} \cdot \boldsymbol{y}}{\|\hat{\boldsymbol{y}}\|\|\boldsymbol{y}\|}\right); \tag{6}$$

where $\cdot$ denotes the inner product between vectors, $\|\cdot\|$ is the Euclidean norm. In our experiments, the mean, median, tri-mean of all the angular errors, mean of the best 25% and the worst 25% errors are reported.

Comparison to State-of-the-Art Methods

Table 1 compares the proposed C^4 with the state-of-the-art methods in terms of the Mean, Median, Tri-mean, the Best 25% and the Worst 25% of angular errors on two datasets. The proposed method can beat most of color constancy algorithms except the FFCC (Barron and Tsai 2017). On one hand, on the Color Checker dataset, our method significantly outperforms the FFCC on all five metrics, especially 18.18% and 15.10% improvement in the Mean and Worst 25% metrics. On the other hand, on the NUS 8-Camera benchmark, although FFCC outperforms on some metrics, our $\mathrm{C}^4_{\text{SqueezeNet-FC4}}$ is better than FFCC on the mean and Worst 25% metrics. Performance gap on the NUS 8-Camera can be explained by the limited size of scenes (*i.e*. each scene generates 8 images with different cameras) leading to less positive effects of data augmentation in our method. More importantly, the C^4 can consistently beat its direct competitors – its backbone AlexNet-FC^4 and SqueezeNet-FC^4 in all five metrics on both datasets, especially in the more challenging scenes as illustrated in Figure 3. In view of the identical network structure for feature encoding, performance gains can only be explained by the design of the cascaded network structure.

Evaluation on Camera-Agnostic Color Constancy

To verify the robustness of our model against appearance inconsistency due to camera sensitivity, we take two disjoint datasets, one for training and the other for testing. Specifically, we conduct an evaluation on the Color Checker dataset with a model trained on the NUS 8-camera dataset and vice versa, whose results are reported in Table 2. Compared to the state-of-the-art statistical GI (Qian et al. 2019), the C^4 achieves competitive performance and performs better in the Worst 25% metric consistently and significantly. Moreover, our C^4 with different backbone CNNs achieve the best performance again among learning-based illumination estimation in all performance metrics on both datasets, which verifies that our model can mitigate negative effects of imaging patterns across cameras owing to its strong generalization capability via progressive refinement and data argumentation.

Discussion about Loss Combination

In our cascaded structure, the combination of loss functions is something worth discussing. We further discuss the design of our loss function with two strategies: a single multiplication loss and the weighted multiply-accumulate loss, with the three-stage C^4 model.

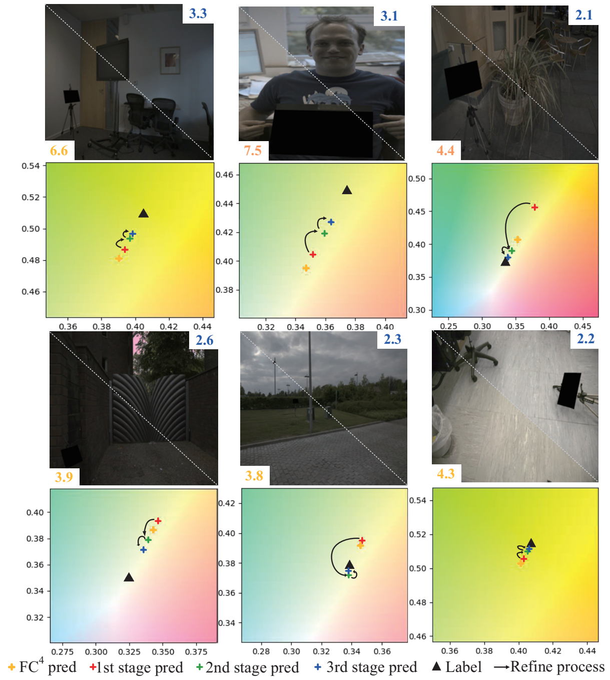


Figure 3: Visualization of harder samples from the Color Checker dataset. In the 1st and 3rd rows, the lower left parts of images are corrected by our detected hypotheses and the other parts are images corrected by the FC^4's predictions. The numbers in the white rectangles of pictures are angle errors (in degrees) between illumination predictions and labels. The 2nd and 4th rows show the trajectories of predictions towards ground truth labels.

- **Single multiplication Loss –** It only penalizes the final illumination prediction. (*e.g.*, in Equation (4), when $L = 3$, weights should be $[w_1, w_2, w_3] = [0, 0, 1]$)
- **Weighted multiply-accumulate loss –** It combines the intermediate illumination prediction from each stage, and penalizes these illumination hypotheses jointly. (*e.g.*, in Equation (4), when $L = 3$, weights satisfying $w_1 \times w_2 \times w_3 \neq 0$)

Table 3 shows comparative results on combined strategies of the loss function. The latter, the weighted multiply-accumulate loss in Equation (4) is superior to its specific case – single multiplication loss, which supports our motivation to design the multiply-accumulate loss to exploit multiple illumination hypotheses. Moreover, among the settings of weights, the equal weight can be slightly better than the remaining, although the improvement is very marginal.

Discussion of Cascade Size

Another key insight of our C^4 is to incrementally improve illumination predictions in a cascaded structure. Performance of such a cascaded structure depends on the size of cascade stages. We demonstrate the validity of our cascaded structure by comparing the performance at varying cascade levels. As shown in Figure 5, angular errors in all metrics of two C^4 variants decrease with cascade level increasing. In particular, the performance increases by a big margin from one-stage C^4 to two-stage variant, while a moderate improvement from two-stage to three-stage, or even to four-stage. However, as the number of cascades continues to increase, the performance does not improve. We suppose that a deeper network makes it harder to fit dramatically increasing size of network parameters. Such a phenomenon encourages a relatively large size of cascade stages for color constancy.

Table 3: Statistics of angular errors (in degrees) obtained by different loss combinations of the three-stage C^4 model on the Color Checker dataset.

w_1	w_2	w_3	Mean	Median	Tri-mean	Best 25%	Worst 25%
Backbone-SqueezeNet-FC^4							
0.00	0.00	1.00	1.48	0.97	1.10	0.32	3.50
0.20	0.30	0.50	1.37	0.92	1.03	0.29	3.26
0.33	0.33	0.34	**1.35**	**0.88**	**0.99**	**0.28**	**3.21**
0.50	0.30	0.20	1.38	0.90	1.00	0.32	3.23
0.70	0.20	0.10	1.37	0.89	1.00	0.31	3.25
Backbone-AlexNet-FC^4							
0.00	0.00	1.00	1.57	1.09	1.22	0.32	3.60
0.20	0.30	0.50	1.52	1.07	1.17	0.32	**3.48**
0.33	0.33	0.34	**1.49**	1.03	**1.13**	**0.29**	3.52
0.50	0.30	0.20	1.50	**1.01**	1.14	0.33	3.50
0.70	0.20	0.10	1.50	1.02	1.14	0.32	3.50

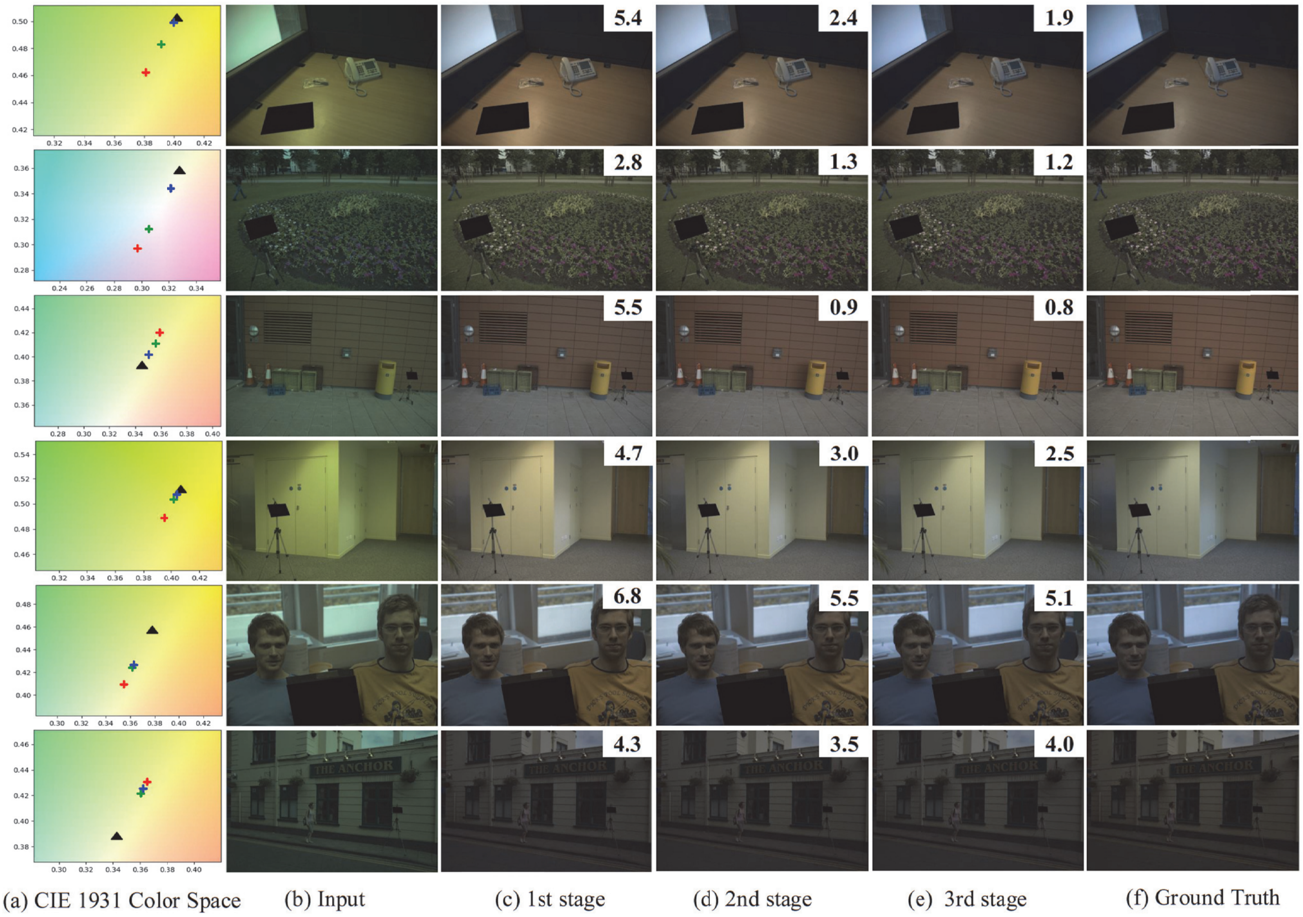
To further illustrate the effectiveness of the introduced cascaded structure, we visualize some examples with intermediate illumination predictions from each stage of the proposed C^4 cascade on the Color Checker dataset in Figure 4. Most corrected images in (c) and (d) are visually closer to ground truth (GT) than those in (b), and we quantitatively measure predictions in the 1st, 2nd and 3rd stages of three-level C^4 model with ground truth of testing samples, $P(1, 2) = 69.72\%$ and $P(2, 3) = 60.21\%$, where $P(l, l+1)$ denotes the ratios of more accurate predictions of the $(l+1)$-th stage in comparison with those of the l-th stage during testing. It further verifies the rationale of the coarse-to-fine cascaded structure.

Table 4: Comparison of network parameters on the Color Checker dataset. The model labeled "1/3p" indicates the backbone network parameters are reduced by one third. "3-stage" means three-stages model. C$^4_{\mathrm{B}}$ and C$^4_{\mathrm{E}}$ mean our proposed network with the model in B) and E) as the backbone respectively. All results in this table are in units of degrees.

Method	Mean	Median	Tri-mean	Best 25%	Worst 25%
A) AlexNet-FC4	1.77	**1.11**	1.29	**0.34**	4.29
B) AlexNet-FC4,1/3p	2.17	1.58	1.71	0.53	4.87
C) C$^4_{\mathrm{B}}$,3 stage	**1.65**	**1.11**	**1.22**	**0.34**	**3.88**
D) SqueezeNet-FC4	1.65	1.18	1.27	0.38	3.78
E) SqueezeNet-FC4,1/3p	1.94	1.40	1.52	0.49	4.31
F) C$^4_{\mathrm{E}}$,3-stage	**1.47**	**0.97**	**1.09**	**0.31**	**3.49**

Evaluation with Comparable Network Parameters

As aforementioned, the performance of such a cascaded structure can be improved with increasing the size of the cascade stages L (when $L <= 4$). However, the number



(a) CIE 1931 Color Space (b) Input (c) 1st stage (d) 2nd stage (e) 3rd stage (f) Ground Truth

Figure 4: Visualization of a number of examples from the Color Checker dataset. (a) In the CIE 1931 color space chromaticity diagram, where the red, green and blue plus sign '+' represent the 1st stage, the 2nd stage and the 3rd stage illumination predictions of our C^4 given input images in (b), the black triangle is the corresponding illumination labels. (c) (d) and (e) are corrected images by intermediate and final illumination hypotheses spotted by the C^4. The angle errors (in degrees) between illumination predictions and labels are highlighted in the white rectangles in the top-right of images.

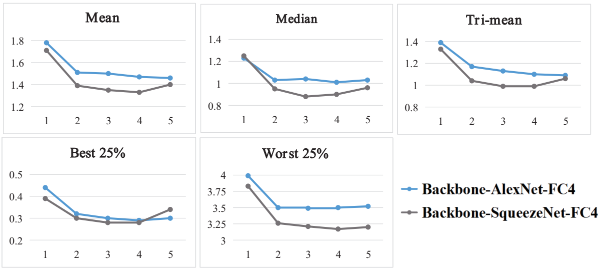


Figure 5: Evaluation on effect of cascade size on the Color Checker dataset. The x axis: the length of the cascade, while the y axis: the angular error (in degrees).

of network parameters is proportional to the size of L. To explore the real source of this improvement, we compress our backbone (*i.e.*, AlexNet-FC4 and SqueezeNet-FC4) by decreasing the size of convolution kernels in every convolutional layer. As shown in Table 4, network parameters in method B) and E) are only one-third of those in original backbone method A) and D) after compressing. When using compressed backbone networks, we get new cascade models (*i.e.* three-stage C^4 methods C) and F)), whose sizes of network parameters are comparable to original FC^4 models in A) and D). Results in Table 4 show that superior performance can be achieved by method C) and F) in comparison with method A) and D), which reveal that performance gain of our C^4 can be credited to the cascade network structure rather than the size of network parameters.

Conclusion

This paper designs a cascade of convolutional neural networks for color constancy, which consistently achieves the best performance for more challenging samples (on the Worst 25% metric) and more robust performance under the camera-agnostic setting. Experiment results favor for a relatively larger cascade size and verify the boosting benefits of combining multiple illumination hypotheses and the coarse-to-fine refinement.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant No.: 61771201, 61902131), the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183), the Fundamental Research Funds for the Central Universities (Grant No.: D2193130), and the SCUT Program (Grant No.: D6192110).

References

Barnard, K.; Cardei, V.; and Funt, B. 2002. A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data. *IEEE transactions on Image Processing*.

Barnard, K. 2000. Improvements to gamut mapping colour constancy algorithms. In *European conference on computer vision*.

Barron, J. T., and Tsai, Y.-T. 2017. Fast fourier color constancy. In *Computer Vision and Pattern Recognition*.

Barron, J. T. 2015. Convolutional color constancy. In *International Conference on Computer Vision*.

Bianco, S.; Cusano, C.; and Schettini, R. 2017. Single and multiple illuminant estimation using convolutional neural networks. *IEEE Transactions on Image Processing*.

Brainard, D. H., and Wandell, B. A. 1986. Analysis of the retinex theory of color vision. *JOSA A*.

Buchsbaum, G. 1980. A spatial processor model for object colour perception. *Journal of the Franklin institute*.

Cardei, V. C., and Funt, B. 1999. Committee-based color constancy. In *Color and Imaging Conference*.

Chakrabarti, A.; Hirakawa, K.; and Zickler, T. 2011. Color constancy with spatio-spectral statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chakrabarti, A. 2015. Color constancy by learning to predict chromaticity from luminance. In *Advances in Neural Information Processing Systems*.

Chen, K.; Jia, K.; Huttunen, H.; Matas, J.; and Kämäräinen, J.-K. 2019. Cumulative attribute space regression for head pose estimation and color constancy. *Pattern Recognition*.

Cheng, D.; Price, B.; Cohen, S.; and Brown, M. S. 2015. Effective learning-based illuminant estimation using simple features. In *Computer Vision and Pattern Recognition*.

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*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*.

Finlayson, G. D., and Trezzi, E. 2004. Shades of gray and colour constancy. In *Color and Imaging Conference*.

Finlayson, G. D. 2013. Corrected-moment illuminant estimation. In *International Conference on Computer Vision*.

Funt, B., and Xiong, W. 2004. Estimating illumination chromaticity via support vector regression. In *Color and Imaging Conference*.

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*.

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

Gijsenij, A., and Gevers, T. 2010. Color constancy using natural image statistics and scene semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hordley, S. D., and Finlayson, G. D. 2004. Re-evaluating colour constancy algorithms. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004.*

Hu, Y.; Wang, B.; and Lin, S. 2017. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Computer Vision and Pattern Recognition*.

Joze, H. R. V., and Drew, M. S. 2013. Exemplar-based color constancy and multiple illumination. *IEEE transactions on pattern analysis and machine intelligence*.

Joze, H. R. V.; Drew, M. S.; Finlayson, G. D.; and Rey, P. A. T. 2012. The role of bright pixels in illumination estimation. In *Color and Imaging Conference*.

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

Qian, Y.; Chen, K.; Kämäräinen, J.-K.; Nikkanen, J.; and Matas, J. 2016. Deep structured-output regression learning for computational color constancy. In *2016 23rd International Conference on Pattern Recognition*.

Qian, Y.; Chen, K.; Nikkanen, J.; Kamarainen, J.-K.; and Matas, J. 2017. Recurrent color constancy. In *International Conference on Computer Vision*.

Qian, Y.; Kamarainen, J.-K.; Nikkanen, J.; and Matas, J. 2019. On finding gray pixels. In *Computer Vision and Pattern Recognition*.

Schaefer, G.; Hordley, S.; and Finlayson, G. 2005. A combined physical and statistical approach to colour constancy. In *Computer Vision and Pattern Recognition*.

Shi, W.; Loy, C. C.; and Tang, X. 2016. Deep specialized network for illuminant estimation. In *European Conference on Computer Vision*.

Shi, L. 2000. Re-processed version of the gehler color constancy dataset of 568 images. *http://www. cs. sfu. ca/˜color/data/*.

Van De Weijer, J.; Gevers, T.; and Gijsenij, A. 2007. Edge-based color constancy. *IEEE Transactions on image processing*.

von Kries, J. 1902. Chromatic adaption. *Festschrift der Albrecht-Ludwigs-Universität*.

Yang, K.-F.; Gao, S.-B.; and Li, Y.-J. 2015. Efficient illuminant estimation for color constancy using grey pixels. In *Computer Vision and Pattern Recognition*.