

# CLCC: Contrastive Learning for Color Constancy

Yi-Chen Lo\*, Chia-Che Chang\*, Hsuan-Chao Chiu, Yu-Hao Huang,  
Chia-Ping Chen, Yu-Lin Chang, Kevin Jou  
MediaTek Inc., Hsinchu, Taiwan

{yichen.lo, chia-che.chang, Hsuanchao.Chiu, justin-yh.huang,  
chiaping.chen, yulin.chang, kevin.jou}@mediatek.com

## Abstract

In this paper, we present CLCC, a novel contrastive learning framework for color constancy. Contrastive learning has been applied for learning high-quality visual representations for image classification. One key aspect to yield useful representations for image classification is to design illuminant invariant augmentations. However, the illuminant invariant assumption conflicts with the nature of the color constancy task, which aims to estimate the illuminant given a raw image. Therefore, we construct effective contrastive pairs for learning better illuminant-dependent features via a novel raw-domain color augmentation. On the NUS-8 dataset, our method provides 17.5% relative improvements over a strong baseline, reaching state-of-the-art performance without increasing model complexity. Furthermore, our method achieves competitive performance on the Gehler dataset with  $3\times$  fewer parameters compared to top-ranking deep learning methods. More importantly, we show that our model is more robust to different scenes under close proximity of illuminants, significantly reducing 28.7% worst-case error in data-sparse regions. Our code is available at <https://github.com/howardyclo/clcc-cvpr21>.

## 1. Introduction

The human visual system can perceive the same canonical color of an object even under different illuminants. This feature can be mimicked by computational color constancy, an essential task in the camera pipeline that processes raw sensor signals to sRGB images. Conventional methods [10, 20, 21, 40, 60] utilize statistical properties of the scene to cope with this ill-posed problem, such as the most widely used gray world assumption. Such statistical methods, however, often fail where their assumptions are violated in complex scenes.

\* Indicates equal contribution.

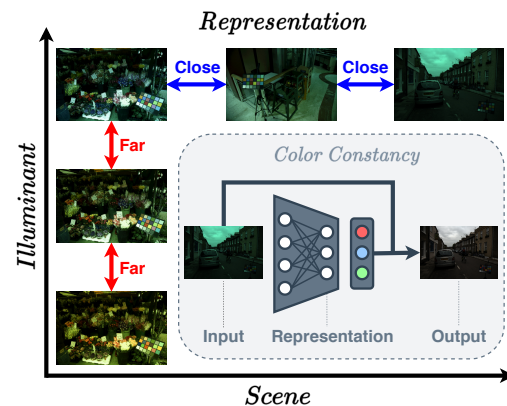


Figure 1: Our main idea of CLCC: The scene-invariant, illuminant-dependent representation of the same scene under different illuminants should be far from each other, while different scenes under the same illuminant should be close to each other.

Until recently, deep learning based methods [31, 48, 66, 67] have been applied to the color constancy problem and achieve considerable quality improvements on challenging scenes. Yet, this ill-posed and sensor-dependent task still suffers from the difficulty of collecting massive paired data for supervised training.

When learning with insufficient training data, a common issue frequently encountered is the possibility of learning spurious correlations [62] or undesirable biases from data [59]: misleading features that work for most training samples but do not always hold in general. For instance, previous research has shown that a deep object-recognition model may rely on the spuriously correlated background instead of the foreground object to make predictions [65] or be biased towards object textures instead of shapes [24]. In the case of color constancy, outdoor scenes often have higher correlations with high color temperature illuminants than indoor scenes. Thus, deep learning models may focus on scene related features instead of illuminant related features.

This leads to a decision behavior that tends to predict high color temperature illuminants for outdoor scenes, but suffers high error on outdoor scenes under low color temperature illuminants. This problem becomes worse when the sparsity of data increases.

To avoid learning such spurious correlations, one may seek to regularize deep learning models to learn scene-invariant, illuminant-dependent representations. As illustrated in Fig.1, in contrast to image classification problem, the representation of the same scene under different illuminants should be *far* from each other. On the contrary, the representation of different scenes under the same illuminant should be *close* to each other. Therefore, we propose to learn such desired representations by contrastive learning [13, 27, 30], a framework that learns general and robust representations by comparing similar and dissimilar samples.

However, conventional self-supervised contrastive learning often generates easy or trivial contrastive pairs that are not very useful for learning generalized feature representations [37]. To address this issue, a recent work [13] has demonstrated that strong data augmentation is crucial for conducting successful contrastive learning.

Nevertheless, previous data augmentations that have been shown effective for image classification may not be suitable for color constancy. Here we illustrate some of them. First, most previous data augmentations in contrastive learning are designed for high-level vision tasks (e.g., object recognition) and seek illuminant invariant features, which can be detrimental for color constancy. For example, color dropping converts an sRGB image to a gray-scale one, making the color constancy task even more difficult. Moreover, the color constancy task works best in the linear color space where the linear relationship to scene radiance is preserved. This prevents from using non-linear color jittering augmentations, e.g., contrast, saturation, and hue.

To this end, we propose *CLCC: Contrastive Learning for Color Constancy*, a novel color constancy framework with contrastive learning. For the purpose of color constancy, effective positive and negative pairs are constructed by exploiting the label information, while novel color augmentations are designed based on color domain knowledge [2, 49, 36].

Built upon a previous state-of-the-art [31], CLCC provides additional 17.5% improvements (mean angular error decreases from 2.23 to 1.84) on a public benchmark dataset [14], achieving state-of-the-art results without increasing model complexity. Besides accuracy improvement, our method also allows deep learning models to effectively acquire robust and generalized representations even when learning from small training datasets.

**Contribution** We introduce CLCC, a fully supervised contrastive learning framework for the task of color constancy. By leveraging label information, CLCC generates more di-

verse and harder contrastive pairs to effectively learn feature representations aiming for better quality and robustness. A novel color augmentation method that incorporates color domain knowledge is proposed. We improve the previous state-of-the-art deep color constancy model without increasing model complexity. CLCC encourages learning illuminant-dependent features rather than spurious scene content features irrelevant for color constancy, making our model more robust and generalized, especially in data-sparse regions.

## 2. Related Work

### 2.1. Contrastive learning

Contrastive learning is a framework that learns general and robust feature representations by comparing similar and dissimilar pairs. Inspired from noise contrastive estimation (NCE) and N-pair loss [26, 45, 55], remarkable improvements on image classification are obtained in several recent works [13, 27, 30, 43, 57, 61, 63]. Particularly, a mutual information based contrastive loss, InfoNCE [61] has become a popular choice for contrastive learning (see [44, 47] for more discussion). Furthermore, recent works [3, 7, 16, 29, 37, 58] have shown that leveraging supervised labels not only improves learning efficiency by alleviating sampling bias (and hence reducing the need for large batch size training) but also improves generalization by learning task-relevant features.

### 2.2. Data augmentation

Data augmentations such as random cropping, flipping, and rotation have been widely used in classification [28, 51], object detection [42], and semantic segmentation [12] to improve model quality. Various works rely on manually designed augmentations to reach their best results [13, 52]. To ease such efforts, strategy search [17, 18] or data synthesis [46, 68] have been used to improve data quality and diversity. However, popular data augmentation strategies for image recognition [13, 17, 34, 50] (e.g., color channel dropping, color channel swapping, HSV jittering) may not be suitable for the color constancy task. Thus, we incorporate color domain knowledge [2, 36, 49] to design data augmentation suitable for contrastive learning on color constancy.

### 2.3. Color constancy

Color constancy is a fundamental low-level computer vision task that has been studied for decades. In general, current research can be divided into learning-free and learning-based approaches. The former ones use color histogram and spatial information to estimate illuminant [10, 20, 21, 40, 60]. Despite the efficiency of these methods, they do not perform well on challenging scenes with ambiguous color pixels. The latter ones adopt data-driven approaches that learn to estimate illuminant from training data [4, 6, 19, 22, 32]. These learning-based approaches outperform learning-free

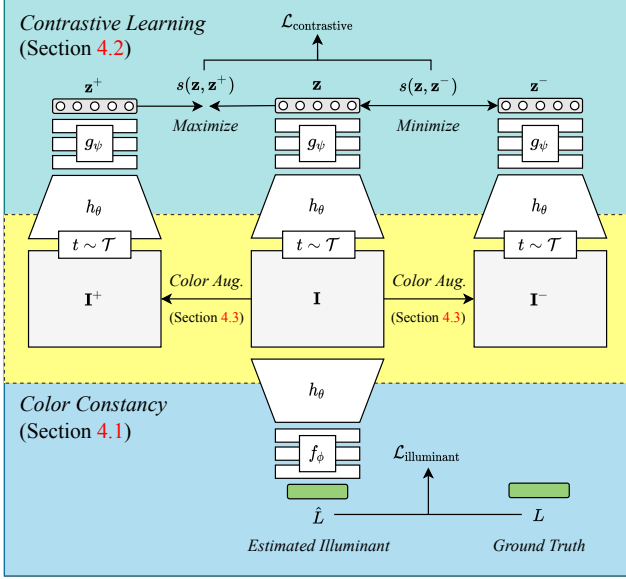


Figure 2: An overview of our CLCC: Besides the main color constancy task, we propose to incorporate contrastive learning to learn generalized and illuminant-dependent feature representations.

methods and have become popular in both academic and industry fields. In addition, recent works have shown that features learned from deep neural networks are better than hand-crafted ones [39, 41, 50]. Consequently, deep learning based color constancy research has gradually received more and more attention. Recently, FC4 uses ImageNet-pretrained backbones [31, 39] to prevent over-fitting and estimate illuminant with two additional convolutional layers. RCC-Net [48] uses a convolutional LSTM to extract features in both spatial and temporal domains to estimate illuminants. C4 [67] proposes a cascaded, coarse-to-fine network for color constancy, stacking three SqueezeNets to improve model quality. To mitigate the issue that the learned representation suffers from being sensitive to image content, IGTN [66] introduces metric learning to learn scene-independent illuminant features. From a different perspective, most learning based methods strongly bind to a single sensor’s spectral sensitivity and thus cannot be generalized to other camera sensors without fine-tuning. Several works [1, 33, 64] have attempted to resolve this issue by training on multiple sensors simultaneously. We note that multi-sensor training is out of the scope of this work, hence we do not compare to this line of research.

### 3. Preliminaries

**Image formation model** A raw-RGB image can be viewed as a measurement of scene radiance within a par-

ticular range of spectrum from a camera sensor:

$$\mathbf{I}_{\text{raw}}(\mathbf{x}) = \int_{\omega} R_c(\lambda)S(\mathbf{x}, \lambda)L(\lambda)d\lambda \quad (1)$$

where  $\lambda$  denotes the wavelength,  $\omega \in [380, 720]$  (nm) is the visible spectrum,  $R_c$  is the spectral sensitivities of the sensor’s color channel  $c \in \{r, g, b\}$ . The term  $S(\mathbf{x}, \lambda)$  denotes the scene’s material reflectance at pixel  $\mathbf{x}$  and  $L(\lambda)$  is the illuminant in the scene, assumed to be spatially uniform. Notably,  $\mathbf{I}_{\text{raw}}$  values are linearly proportional to the scene radiance, making color constancy easier to work with.

**Color space conversions** Usually  $\mathbf{I}_{\text{raw}}$  undergoes two color space conversions in the camera pipeline:

$$\mathbf{I}_{\text{sRGB}} = \mathcal{G}_{\text{XYZ} \rightarrow \text{sRGB}}(\mathcal{F}_{\text{raw} \rightarrow \text{XYZ}}(\mathbf{I}_{\text{raw}})) \quad (2)$$

where  $\mathcal{F}(\cdot)$  involves linear operations including white balance and full color correction.  $\mathcal{F}(\cdot)$  maps a sensor-specific raw-RGB to a standard perceptual color space such as CIE XYZ.  $\mathcal{G}(\cdot)$  involves non-linear photo-finishing procedures (e.g., contrast, hue, saturation) and eventually maps XYZ to the sRGB color space (we refer to [35] for a complete overview of camera imaging pipeline).

**White balance and full color correction** Given  $\mathbf{I}_{\text{raw}}$ , white balance (WB) aims to estimate the scene illuminant  $L = [L_r, L_g, L_b]$ , i.e., the color of a neutral material captured with a physical color checker placed in the scene. Knowing that a neutral material equally reflects spectral energy at every wavelength regardless of different illuminants, we can apply a  $3 \times 3$  diagonal matrix  $\mathbf{M}_{\text{WB}}$  with the diagonal entries  $[L_g/L_r, 1, L_g/L_b]$  on  $\mathbf{I}_{\text{raw}}$  to obtain a white-balanced image  $\mathbf{I}_{\text{WB}}$ :

$$\mathbf{I}_{\text{WB}} = \mathbf{I}_{\text{raw}}\mathbf{M}_{\text{WB}} \quad (3)$$

After WB, a neutral material should appear achromatic (i.e., “gray”). Because WB only corrects achromatic colors, a  $3 \times 3$  full color correction matrix  $\mathbf{M}_{\text{CC}}$  is further applied to correct chromatic colors (in practice, those chromatic patches with known CIE XYZ values on color checker). Note that  $\mathbf{M}_{\text{CC}}$  is illuminant-specific due to error introduced by the estimated  $\mathbf{M}_{\text{WB}}$

$$\mathbf{I}_{\text{XYZ}} = \mathbf{I}_{\text{WB}}\mathbf{M}_{\text{CC}} \quad (4)$$

Such  $\mathbf{I}_{\text{XYZ}}$  is sensor-agnostic since the illuminant cast is completely removed for both achromatic and chromatic colors.

### 4. Methodology

We start with our problem formulation and review conventional self-supervised contrastive learning in Section 4.1.

Next, we introduce CLCC, our fully-supervised contrastive learning framework for color constancy in Section 4.2. Finally, we describe our color augmentation for contrastive pair synthesis in Section 4.3. How these sections fit together is illustrated in Fig. 2.

#### 4.1. Formulation

**The learning problem** Our problem setting follows the majority of learning-based color constancy research which only focuses on the white balance step of estimating the illuminant  $L$  from the input raw image  $\mathbf{I}_{\text{raw}}$ :

$$\hat{L} = f_{\phi}(h_{\theta}(\mathbf{I}_{\text{raw}})) \quad (5)$$

where  $h_{\theta}$  is the feature extractor that produces visual representations for  $\mathbf{I}_{\text{raw}}$ ,  $f_{\phi}$  is the illuminant estimation function, and  $\hat{L}$  is the estimated illuminant. Both  $h_{\theta}$  and  $f_{\phi}$  are parameterized by deep neural networks with arbitrary architecture design, where  $\theta$  and  $\phi$  can be trained via back-propagation.

**The learning objectives** The overall learning objective can be decomposed into two parts: (1) illuminant estimation for color constancy and (2) contrastive learning for better representations (as shown in Fig. 2):

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{illuminant}} + \beta \mathcal{L}_{\text{contrastive}} \quad (6)$$

For the illuminant estimation task, we use the commonly used angular error as:

$$\mathcal{L}_{\text{illuminant}} = \arccos\left(\frac{\hat{L} \cdot L}{\|\hat{L}\| \cdot \|L\|}\right) \quad (7)$$

where  $\hat{L}$  is the estimated illuminant and  $L$  is the ground-truth illuminant.

Since the datasets for color constancy are relatively small because it is difficult to collect training data with corresponding ground-truth illuminants. Training a deep learning model with only the supervision  $\mathcal{L}_{\text{illuminant}}$  usually does not generalize well. Therefore, we propose to use contrastive learning, which can help to learn a color constancy model that generalize better even with a small training dataset. Details of the contrastive learning task are described as follows.

**The contrastive learning framework** The proposed CLCC is built upon the recent work SimCLR [13]. Therefore, we discuss self-supervised contrastive learning for color constancy in this section, and then elaborate on our extended fully-supervised contrastive learning in the next section. The essential building blocks of contrastive learning are illustrated here:

- A stochastic data augmentation  $t(\cdot) \sim \mathcal{T}$  that augments a sample image  $\mathbf{I}$  to a different *view*  $t(\mathbf{I})$ . Note that  $t(\cdot)$  is required to be *label-preserving*, meaning that  $\mathbf{I}$  and  $t(\mathbf{I})$  still share the same ground-truth illuminant  $L$ .

- A feature extraction function  $h_{\theta}$  that extracts the *representation* of  $t(\mathbf{I})$ .  $h_{\theta}$  is further used for downstream color constancy task as defined in the Eq. (5).
- A feature projection function  $g_{\psi}$  that maps the representation  $h_{\theta}(t(\mathbf{I}))$  to the *projection*  $\mathbf{z}$  that lies on a unit hypersphere.  $g_{\psi}$  is typically only required when learning representations and thrown away once the learning is finished.
- A similarity metric function  $s(\cdot)$  that measures the similarity between latent projections  $(\mathbf{z}_i, \mathbf{z}_j)$ .
- Contrastive pair formulation: *anchor*  $\mathbf{I}$ , *positive*  $\mathbf{I}^+$  and *negative*  $\mathbf{I}^-$  samples jointly compose the positive pair  $(\mathbf{I}, \mathbf{I}^+)$  and the negative pair  $(\mathbf{I}, \mathbf{I}^-)$  for contrastive learning. For the color constancy task, a positive pair should share the same illuminant label  $L$ , while a negative pair should have different ones.
- A contrastive loss function  $\mathcal{L}_{\text{contrastive}}$  that aims to maximize the similarity between the projection of the positive pair  $(\mathbf{z}, \mathbf{z}^+)$  and minimize the similarity between that of the negative pair  $(\mathbf{z}, \mathbf{z}^-)$  in the latent projection space.

**Self-supervised contrastive learning** Given two random training images  $\mathbf{I}_i$  and  $\mathbf{I}_j$  with different scene content, one can naively form a positive contrastive pair with two randomly augmented views of the same image  $(t(\mathbf{I}_i), t'(\mathbf{I}_i^+))$ , and a negative contrastive pair with views of two different images  $(t(\mathbf{I}_i), t'(\mathbf{I}_j^-))$ .

Such naive formulation introduces two potential drawbacks. One is the *sampling bias*, the potential to sample a false negative pair that shares very similar illuminants (i.e.,  $L_i \simeq L_j$ ). The other is the *lack of hardness*, the fact that the positive  $t(\mathbf{I}_i^+)$  derived from the same image as the anchor  $t(\mathbf{I}_i)$  could share similar scene content. This alone suffices to let neural networks easily distinguish from negative  $t'(\mathbf{I}_j^-)$  with apparently different scene content. Hence, as suggested by [13], one should seek strong data augmentations to regularize such learning shortcut.

To alleviate sampling bias and increase the hardness of contrastive pairs, we propose to leverage label information, extending self-supervised contrastive learning into fully-supervised contrastive learning, where the essential data augmentation is specifically designed to be label-preserving for color constancy task.

#### 4.2. CLCC: Contrastive learning for color constancy

We now describe our realization of each component in the proposed fully-supervised contrastive learning framework, as depicted in Fig. 3

**Contrastive pair formulation** Here, we define  $\mathbf{I}_{\text{XA}}$  as a linear raw-RGB image captured in the scene X under the illu-

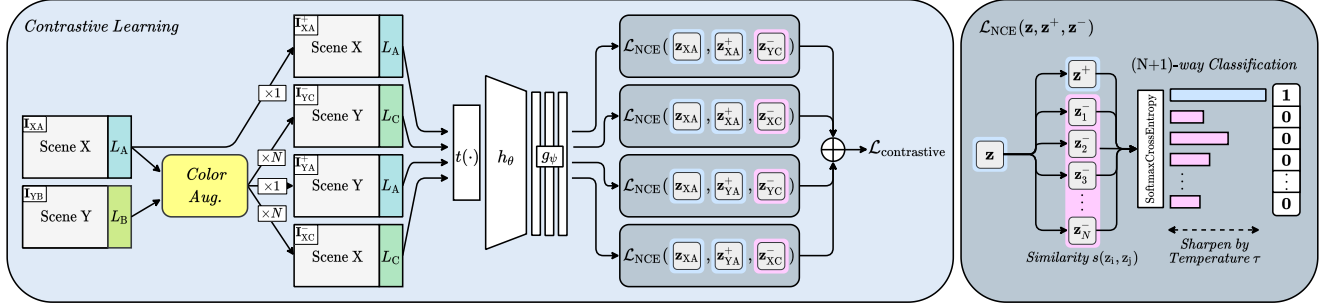


Figure 3: The proposed formation for contrastive pairs and color augmentation.

inant  $L_A$ . Let us recapitulate our definition that a positive pair should share an identical illuminant while a negative pair should not. Therefore, given two randomly sampled training images  $\mathbf{I}_{AX}$  and  $\mathbf{I}_{BY}$ , we construct our contrastive pairs as follows:

- An easy positive pair  $(t(\mathbf{I}_{XA}), t'(\mathbf{I}_{XA}^+))$ —with an identical scene X and illuminant  $L_A$ .
- An easy negative pair  $(t(\mathbf{I}_{XA}), t'(\mathbf{I}_{YC}^-))$ —with different scenes (X, Y) and different illuminants ( $L_A, L_C$ ).
- A hard positive pair  $(t(\mathbf{I}_{XA}), t'(\mathbf{I}_{YA}^+))$ —with different scenes (X, Y) but an identical illuminant  $L_A$ .
- A hard negative pair  $(t(\mathbf{I}_{XA}), t'(\mathbf{I}_{XC}^-))$ —with an identical scene X but different illuminants ( $L_A, L_C$ ).

$\mathbf{I}_{YC}$ ,  $\mathbf{I}_{YA}$  and  $\mathbf{I}_{XC}$  are synthesized by replacing one scene’s illuminant to another. Note that we define the novel illuminant  $L_C$  as the interpolation or extrapolation between  $L_A$  and  $L_B$ , thus we do not need a redundant hard negative sample  $\mathbf{I}_{XB}$ . More details are explained in Section 4.3.  $t$  is a stochastic perturbation-based, illuminant-preserving data augmentation composed by *random intensity*, *random shot noise*, and *random Gaussian noise*.

**Similarity metric and contrastive loss function** Once the contrastive pairs are defined in the image space, we use  $h_\theta$  and  $g_\psi$  to encode those views  $t(\cdot)$  to the latent projection space  $\mathbf{z}$ . Our contrastive loss can be computed as the sum of InfoNCE losses for properly elaborated contrastive pairs:

$$\begin{aligned} \mathcal{L}_{\text{contrastive}} = & \mathcal{L}_{\text{NCE}}(\mathbf{z}_{XA}, \mathbf{z}_{XA}^+, \mathbf{z}_{YC}^-) \\ & + \mathcal{L}_{\text{NCE}}(\mathbf{z}_{XA}, \mathbf{z}_{XA}^+, \mathbf{z}_{XC}^-) \\ & + \mathcal{L}_{\text{NCE}}(\mathbf{z}_{XA}, \mathbf{z}_{YA}^+, \mathbf{z}_{YC}^-) \\ & + \mathcal{L}_{\text{NCE}}(\mathbf{z}_{XA}, \mathbf{z}_{YA}^+, \mathbf{z}_{XC}^-) \end{aligned} \quad (8)$$

The InfoNCE loss  $\mathcal{L}_{\text{NCE}}$  can be computed as:

$$\mathcal{L}_{\text{NCE}} = -\log \left[ \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{n=1}^N \exp(s^-/\tau)} \right] \quad (9)$$

where  $s^+$  and  $s^-$  are the cosine similarity scores of positive and negative pairs respectively:

$$\begin{aligned} s^+ &= s(\mathbf{z}, \mathbf{z}^+) \\ s^- &= s(\mathbf{z}, \mathbf{z}^-) \end{aligned} \quad (10)$$

Equation (9) could be viewed as performing a  $(N + 1)$ -way classification realized by cross-entropy loss with  $N$  negative pairs and 1 positive pair.  $\tau$  is the temperature scaling factor.

### 4.3. Raw-domain Color Augmentation

The goal of our proposed color augmentation is to synthesize more diverse and harder positive and negative samples by manipulating illuminants such that the color constancy solution space is better constrained. As shown in Fig. 4, for example, given two randomly sampled  $(\mathbf{I}_{XA}, L_A)$ , and  $(\mathbf{I}_{YB}, L_B)$  from training data, we go through the following procedure to synthesize  $\mathbf{I}_{YC}$ ,  $\mathbf{I}_{YA}$  and  $\mathbf{I}_{XC}$ , as defined in Section 4.2.

**Color checker detection** We extract 24 linear-row RGB colors  $\mathbf{C}_A \in \mathbb{R}^{24 \times 3}$  and  $\mathbf{C}_B \in \mathbb{R}^{24 \times 3}$  of the color checker from  $\mathbf{I}_{XA}$  and  $\mathbf{I}_{YB}$  respectively using the off-the-shelf color checker detector.

**Color transformation matrix** Given  $\mathbf{C}_A$  and  $\mathbf{C}_B$ , we can solve a linear mapping  $\mathbf{M}_{AB} \in \mathbb{R}^{3 \times 3}$  that transform  $\mathbf{C}_A$  to  $\mathbf{C}_B$  by any standard least-square method. The inverse mapping  $\mathbf{M}_{BA}$  can be derived by solving the  $\mathbf{M}_{AB}^{-1}$ . Accordingly, we can augment  $\mathbf{I}_{XB}$  and  $\mathbf{I}_{YA}$  as:

$$\begin{aligned} \mathbf{I}_{XB} &= \mathbf{I}_{XA} \mathbf{M}_{AB} \\ \mathbf{I}_{YA} &= \mathbf{I}_{YB} \mathbf{M}_{BA} \end{aligned} \quad (11)$$

**Novel illuminant synthesis** The above augmentation procedure produces novel samples  $\mathbf{I}_{XB}$  and  $\mathbf{I}_{YA}$ , but using only pre-existing illuminants  $L_A$  and  $L_B$  from the training data. To synthesize a novel sample  $\mathbf{I}_{XC}$  under a novel illuminant  $L_C$  that *does not exist in the training dataset*, we can synthesize  $\mathbf{C}_C$  by channel-wise *interpolating* or *extrapolating*



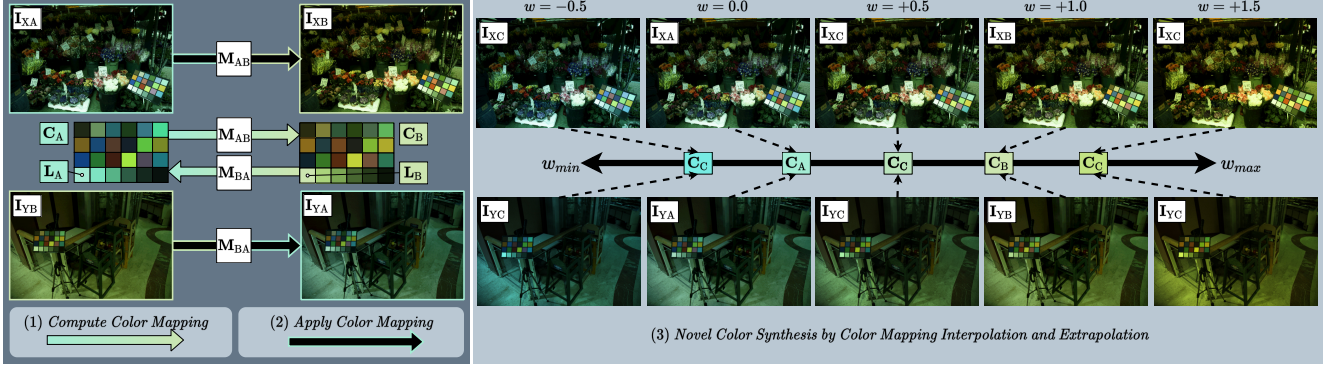


Figure 4: An illustration of our proposed color augmentation. The left hand side shows the generation of positive/negative samples by swapping pre-existing illuminants from a pair of images via estimated color mapping matrices  $M_{AB}$  and  $M_{BA}$ . The right hand side shows the augmented samples with novel illuminants via interpolation ( $w = +0.5$ ) and extrapolation ( $w = -1.5$  and  $w = +1.5$ ) using the detected color checkers  $C_A$  and  $C_B$ .

from the existing  $C_A$  and  $C_B$  as:

$$C_C = (1 - w)C_A + wC_B \quad (12)$$

where  $w$  can be randomly sampled from a uniform distribution of an appropriate range  $[w_{min}, w_{max}]$ . Note that  $w$  should not be close to zero in avoidance of yielding a false negative sample  $I_{XC} = I_{XA}$  for contrastive learning.

To more realistically synthesize  $I_{XC}$  (i.e., more accurate on chromatic colors), we need the full color transformation matrix  $M_{AC}$  that maps  $I_{XA}$  to  $I_{XC}$ :

$$\begin{aligned} I_{XC} &= I_{XA}M_{AC} \\ I_{YC} &= I_{YB}M_{BC} \end{aligned} \quad (13)$$

where  $M_{AC}$  can be efficiently computed from the identity matrix  $\mathbb{1}$  and  $M_{AB}$  without solving least-squares as:

$$\begin{aligned} M_{AC} &= (1 - w)\mathbb{1} + wM_{AB} \\ M_{BC} &= w\mathbb{1} + (1 - w)M_{BA} \end{aligned} \quad (14)$$

Equation (14) can be derived from Eq.(12) and Eq.(13).

**From full color mapping to neutral color mapping** Our synthesis method could be limited by the performance of color checker detection. When the color checker detection is not successful, the full colors  $C_A$  and  $C_B$  could be reduced to the neutral ones  $L_A$  and  $L_B$ , meaning that the color transformation matrix  $M_{AB}$  is reduced from a full matrix to a diagonal matrix. This is also equivalent to first perform WB on  $I_A$  with  $L_A$ , and subsequently perform an inverse WB with  $L_B$ .

We provide the ablation study of this simplified version in our experiment, where we term the full color mapping as *Full-Aug* and the simplified neutral color mapping as *WB-Aug*. We show that even though chromatic colors cannot be correctly mapped, *WB-Aug* could still obtain performance improvement over the baseline.

## 5. Experiment

### 5.1. Network training

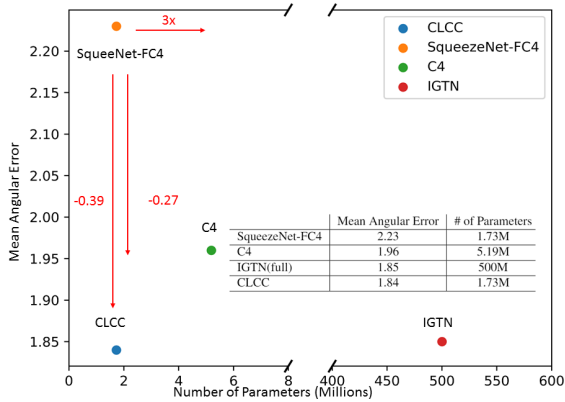
Following FC4 [31], we use ImageNet-pretrained SqueezeNet as the backbone and add a non-linear projection head with three-layer MLP with 512 hidden units for contrastive learning. Note that the projection head is thrown away once the learning is finished. We use Adam [38] optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . The learning rate is 0.0003 and batch size is 16. We use dropout [56] with probability of 0.5 and  $L_2$  weight decay of 0.000057 for regularization. The loss weights for illuminant estimation and contrastive learning heads ( $\lambda, \beta$ ) is (0.1, 1.0) for the first 5000 epochs, (1.0, 0.1) for the rest 5000 epochs in learning objective (6). The number of negative samples  $N$  is 12 and the temperature scaling factor  $\tau$  is 0.87 for InfoNCE loss (9). Note that we do not train our illuminant estimation head with contrastive pairs. They are only used for training the contrastive learning head as depicted in Fig. 2.

### 5.2. Data augmentation

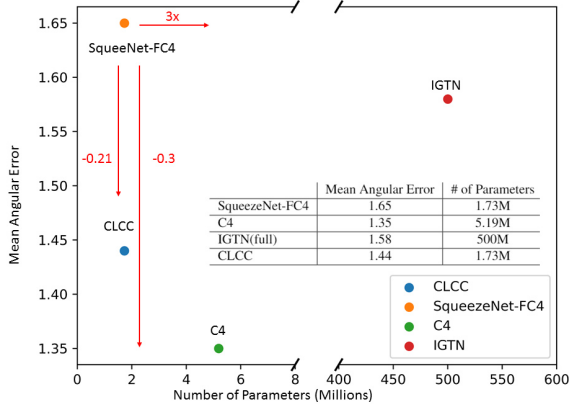
We follow the default data augmentations used by FC4 with several differences. We resize the crop to  $256 \times 256$  to speed-up training. For perturbation-based augmentations in contrastive learning, the range of intensity gain is  $[0.8, 1.2]$ , and the ranges of standard deviation of Gaussian noise and shot noise are  $[0, 0.04]$  and  $[0.02, 0.06]$  for  $[0, 1]$ -normalized images respectively. The  $(w_{min}, w_{max})$  for novel color synthesis (12) are  $(-5.0, -0.3)$  and  $(+0.3, +5.0)$ .

### 5.3. Dataset and evaluation metric

There are two standard public datasets for color constancy task: the reprocessed [53] Color Checker Dataset [23] (termed as the Gehler dataset in this paper) and the NUS-8 Dataset [14]. The Gehler dataset has 568 linear raw-RGB



(a) On the NUS-8 dataset, CLCC achieves the best results and the most light-weighted model of all comparable methods.



(b) On the Gehler dataset, without increasing model complexity, CLCC improves SqueezeNet-FC4 to achieve comparable results.

Figure 5: Model complexity versus mean angular error. CLCC improves SqueezeNet-FC4 by 0.39 (17.5%) on the NUS-8 dataset as the state-of-the-art, and by 0.21 (12.5%) on the Gehler dataset as a comparable method.

images captured by 2 cameras and the NUS-8 dataset has 1736 linear raw-RGB images captured by 8 cameras. In color constancy studies, three-fold cross validation is widely used for both datasets. Several standard metrics are reported in terms of angular error in degrees: mean, median, tri-mean of all the errors, mean of the lowest 25% of errors, and mean of the highest 25% of errors.

## 5.4. Evaluation

**Quantitative evaluation** Following the evaluation protocol, we perform three-fold cross validation on the NUS-8 and the Gehler datasets. We compare our performance with

	Mean	Median	Tri.	Best-25%	Worst-25%
White-Patch [9]	10.62	10.58	10.49	1.86	19.45
Edge-based Gamut [4]	8.43	7.05	7.37	2.41	16.08
Pixel-based Gamut [4]	7.70	6.71	6.90	2.51	14.05
Intersection-based Gamut [4]	7.20	5.96	6.28	2.20	13.61
Gray-World [10]	4.14	3.20	3.39	0.90	9.00
Bayesian [23]	3.67	2.73	2.91	0.82	8.21
NIS [25]	3.71	2.60	2.84	0.79	8.47
Shades-of-Gray [20]	3.40	2.57	2.73	0.77	7.41
1st-order Gray-Edge [60]	3.20	2.22	2.43	0.72	7.36
2nd-order Gray-Edge [60]	3.20	2.26	2.44	0.75	7.27
Spatio-spectral (GenPrior) [11]	2.96	2.33	2.47	0.80	6.18
Corrected-Moment (Edge) [19]	3.03	2.11	2.25	0.68	7.08
Corrected-Moment (Color) [19]	3.05	1.90	2.13	0.65	7.41
Cheng et al. [14]	2.92	2.04	2.24	0.62	6.61
CCC (dist+ext) [5]	2.38	1.48	1.69	0.45	5.85
Regression TreeTree [15]	2.36	1.59	1.74	0.49	5.54
DS-Net (HypNet + SelNet) [54]	2.24	1.46	1.68	0.48	6.08
FFCC-4 channels [6]	1.99	<b>1.31</b>	<b>1.43</b>	<b>0.35</b>	4.75
AlexNet-FC4 [31]	2.12	1.53	1.67	0.48	4.78
SqueezeNet-FC4 [31]	2.23	1.57	1.72	0.47	5.15
IGTN (vanilla triplet loss) [66]	2.02	1.36	-	0.45	4.70
IGTN (no triplet loss) [66]	2.28	1.64	-	0.51	5.20
IGTN (no learnable histogram) [66]	2.15	1.52	-	0.47	5.28
IGTN (full) [66]	1.85	<b>1.24</b>	-	<b>0.36</b>	4.58
C4 [67]	1.96	1.42	1.53	0.48	<b>4.40</b>
CLCC w/ Full-Aug	<b>1.84</b>	<b>1.31</b>	<b>1.42</b>	0.41	<b>4.20</b>

Table 1: Angular error of various methods on the NUS-8 dataset. CLCC gets the best results on the mean tri-mean and worst-25% metrics, and comparable results on the others. Notably, although IGTN gets the best result on the median metric, its model complexity is the largest.

	Mean	Median	Tri.	Best-25%	Worst-25%	Extra data
Gray World [10]	6.36	6.28	6.28	2.33	10.58	
General Gray World [60]	4.66	3.48	3.81	1.00	10.58	
White Patch [9]	7.55	5.68	6.35	1.45	16.12	
Shades-of-Gray [20]	4.93	4.01	4.23	1.14	10.20	
Spatio-spectral (GenPrior) [11]	3.59	2.96	3.10	0.95	7.61	
Cheng et al. [14]	3.52	2.14	2.47	0.50	8.74	
NIS [25]	4.19	3.13	3.45	1.00	9.22	
Corrected-Moment (Edge) [19]	3.12	2.38	-	0.90	6.46	
Corrected-Moment (Color) [19]	2.96	2.15	-	0.64	6.69	
Exemplar [32]	3.10	2.30	-	-	-	
Regression Tree [15]	2.42	1.65	1.75	0.38	5.87	
CNN [8]	2.36	1.95	-	-	-	
CCC (dist+ext) [5]	1.95	1.38	1.22	0.35	4.76	
DS-Net (HypNet+SelNet) [54]	1.90	1.12	1.33	0.31	4.84	
FFCC-4 channels [6]	1.78	0.96	1.14	0.29	4.29	
FFCC-2 channels [6]	1.67	0.96	1.13	0.26	4.23	+S
FFCC-2 channels [6]	1.65	<b>0.86</b>	1.07	0.24	4.44	+M
FFCC-2 channels [6]	1.61	<b>0.86</b>	<b>1.02</b>	<b>0.23</b>	4.27	+S+M
AlexNet-FC4 [31]	1.77	1.11	1.29	0.34	4.29	
SqueezeNet-FC4 [31]	1.65	1.18	1.27	0.38	3.78	
IGTN (vanilla triplet loss) [66]	1.73	1.09	-	0.31	4.25	
IGTN (no triplet loss) [66]	1.78	1.13	-	0.34	4.31	
IGTN (no learnable histogram) [66]	1.85	1.10	-	0.31	4.91	
IGTN (full) [66]	1.58	0.92	-	0.28	3.70	
C4 [67]	<b>1.35</b>	<b>0.88</b>	<b>0.99</b>	0.28	<b>3.21</b>	
CLCC w/ Full-Aug	<b>1.44</b>	0.92	1.04	<b>0.27</b>	<b>3.48</b>	

Table 2: Angular error of various methods on the Gehler dataset. The use of semantic data or meta-data are denoted by “S” or “M”. The result shows that SqueezeNet-FC4 plugging in our approach, which keeps the same model complexity and without meta data can achieve comparable performance.

previous state-of-the-art approaches. As shown in Fig. 5a, the proposed CLCC is able to achieve state-of-the-art mean angular error on the NUS-8 dataset, 17.5% improvements compared to FC4 with similar model size. Other competitive methods, such as C4 and IGTN, use much more model parameters (3× and more than 200×) but give worse mean angular error. Table 1 shows comprehensive performance

comparisons with recent methods on the NUS-8 dataset [14]. Our CLCC provides significant improvements over the baseline network SqueezeNet-FC4 across all scoring metrics and reach the best mean metric, as well as the best worst-25% metric. This indicates that the proposed fully-supervised contrastive learning not only improves the overall performance when there is no massive training data, but also improves robustness via effective constrastive pairs constructions. For the Gehler dataset, as shown in Fig. 5b, our CLCC stays competitive with less than 0.1 performance gap behind the best performing approach C4 [67], whose model size is  $3\times$  larger. Table 2 shows detailed performance of state-of-the-art methods on the Gehler dataset. It is shown that methods achieving better scores than CLCC either require substantially more complexity (C4), or utilize supplemental data (FFCC). C4 has three times more parameters which may facilitate remembering more sensor features than ours. FFCC needs meta-data from camera to reach the best median metric. If no auxiliary data is used, CLCC performs better than FFCC-4 channels on all metrics.

**Ablation for color augmentation** Recap that our proposed color augmentation methods for contrastive learning includes Full-Aug and WB-Aug mentioned in Section 4.3. As shown in Table 3, even when the color checker is not successfully detected for full color mapping (Full-Aug), the reduced neutral color mapping (WB-Aug) is still able to significantly decrease the mean angular error from 2.23 to 1.93 and the worst-case error from 5.15 to 4.30 over the SqueezeNet-FC4 baseline, which are substantial relative improvement 13.5% and 16.5% respectively. Furthermore, when Full-Aug is considered, the mean angular error can be decreased from 1.93 to 1.84 with an additional relative improvement 5.1%. This shows that correctly mapped chromatic colors for synthesizing contrastive pairs can improve the quality of contrastive learning, resulting a improved model.

	Mean	Median	Best-25%	Worst-25%
FC4 [31]	2.23	1.57	0.47	5.15
+ CLCC w/ WB-Aug	1.93	1.38	0.44	4.30
+ CLCC w/ Full-Aug	1.84	1.31	0.41	4.20

Table 3: The results show CLCC is able to improve SqueezeNet-FC4 quality by contrastive learning with two different data augmentations on the NUS-8 dataset.

**Worst-case robustness** In this section, we are also interested in whether CLCC can provide improvements on robustness for worst-cases. To illustrate the robustness in more finegrained level, we propose to evaluate our model under  $K$  grouped data on the Gehler dataset via clustering the illuminant labels with K-means.  $K$  is selected as 5 for example. Each group represents different scene contents under similar

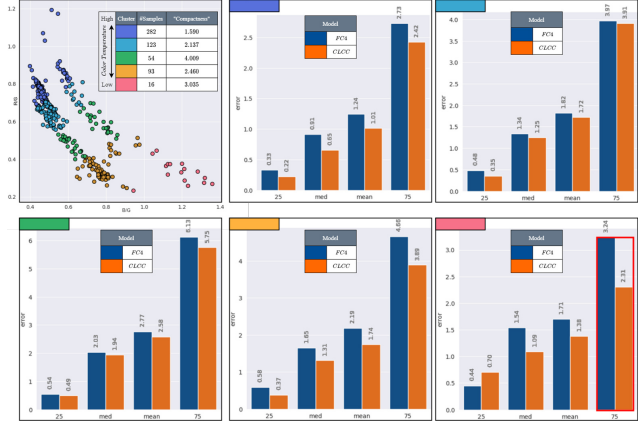


Figure 6: Per-cluster error metrics on the Gehler dataset. We show that CLCC achieves better performance on all clusters, especially the worst-case performance (i.e. worst-25%). Notably, in the sparse data regime (cluster colored with pink that contains only 16 data points), CLCC trades best-case performance (i.e., best-25%) with the worst-case one, leading to better robustness (i.e., lower test error standard deviation).

illuminants. As shown in Fig. 6, CLCC greatly improves on all scoring metrics among all clusters (except for best-25% in pink cluster). Remarkably, we demonstrate that, when the amount of cluster data decreases from higher one (e.g., purple cluster) to lower one (e.g., pink cluster), as shown in the data distribution on top-left side in Fig. 6), the improvement over worse-case performance increases. Especially in the region that suffers from data sparsity (e.g., 16 data points in pink cluster), CLCC significantly reduces the worse-case error from 3.24 to 2.31, which achieves 28.7% relative improvement. This finding supports our contrastive learning design which aims to learn better illuminant-dependent features that are robust and invariant to scene contents.

## 6. Conclusion

In this paper, we present CLCC, a contrastive learning framework for color constancy. Our framework differs from conventional self-supervised contrastive learning on the novel fully-supervised construction of contrastive pairs, driven by our novel color augmentation. We improve considerably over previous strong baseline, achieving state-of-the-art or competitive results on two public benchmark datasets, without additional computational costs. Our design of contrastive pairs allows model to learn better illuminant features that are particularly robust to worse-cases in data sparse regions.



## References

- [1] Mahmoud Afifi and Michael S. Brown. Sensor-independent illumination estimation for DNN models. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019.
- [2] Mahmoud Afifi and Michael S. Brown. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019.
- [3] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019.
- [4] Kobus Barnard. Improvements to gamut mapping colour constancy algorithms. In David Vernon, editor, *Computer Vision - ECCV 2000, 6th European Conference on Computer Vision, Dublin, Ireland, June 26 - July 1, 2000, Proceedings, Part I*, 2000.
- [5] Jonathan T. Barron. Convolutional color constancy. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015.
- [6] Jonathan T. Barron and Yun-Ta Tsai. Fast fourier color constancy. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017.
- [7] Lucas Beyer, Xiaohua Zhai, Avital Oliver, and Alexander Kolesnikov. S4L: self-supervised semi-supervised learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019.
- [8] Simone Bianco, Claudio Cusano, and Raimondo Schettini. Single and multiple illuminant estimation using convolutional neural networks. *IEEE Trans. Image Process.*, 2017.
- [9] David H Brainard and Brian A Wandell. Analysis of the retinex theory of color vision. *JOSA A*, 1986.
- [10] Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 1980.
- [11] Ayan Chakrabarti, Keigo Hirakawa, and Todd E. Zickler. Color constancy with spatio-spectral statistics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.
- [12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, Lecture Notes in Computer Science, 2018.
- [13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. 2020.
- [14] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*, 2014.
- [15] Dongliang Cheng, Brian L. Price, Scott Cohen, and Michael S. Brown. Effective learning-based illuminant estimation using simple features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015.
- [16] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debaised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [17] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 113–123, 2019.
- [18] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, 2020.
- [19] Graham D. Finlayson. Corrected-moment illuminant estimation. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, 2013.
- [20] Graham D. Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *The Twelfth Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications, CIC 2004, Scottsdale, Arizona, USA, November 9-12, 2004*, 2004.
- [21] Brian V. Funt and Lilong Shi. The rehabilitation of maxrgb. In *18th Color and Imaging Conference, CIC 2010, San Antonio, Texas, USA, November 8-12, 2010*. IS&T - The Society for Imaging Science and Technology, 2010.
- [22] Brian V. Funt and Weihua Xiong. Estimating illumination chromaticity via support vector regression. In *The Twelfth Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications, CIC 2004, Scottsdale, Arizona, USA, November 9-12, 2004*, 2004.
- [23] Peter V. Gehler, Carsten Rother, Andrew Blake, Thomas P. Minka, and Toby Sharp. Bayesian color constancy revisited. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.
- [24] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*. OpenReview.net, 2019.
- [25] Arjan Gijsenij and Theo Gevers. Color constancy using natural image statistics and scene semantics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011.
- [26] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and D. Mike Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010*,

- Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, JMLR Proceedings, 2010.
- [27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020.
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.
- [29] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. 2020.
- [30] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.
- [31] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc<sup>4</sup>: Fully convolutional color constancy with confidence-weighted pooling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017.
- [32] Hamid Reza Vaezi Joze and Mark S. Drew. Exemplar-based colour constancy. In Richard Bowden, John P. Collomosse, and Krystian Mikolajczyk, editors, *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, 2012.
- [33] Daniel Hernández Juárez, Sarah Parisot, Benjamin Busam, Ales Leonardis, Gregory G. Slabaugh, and Steven McDonagh. A multi-hypothesis approach to color constancy. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020.
- [34] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 2017.
- [35] Hakki Can Karaimer and M. S. Brown. A software platform for manipulating the camera imaging pipeline. In *ECCV*, 2016.
- [36] Hakki Can Karaimer and Michael S. Brown. Improving color reproduction accuracy on cameras. In *CVPR*, pages 6440–6449. IEEE Computer Society, 2018.
- [37] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. 2020.
- [38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. 2012.
- [40] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61.
- [41] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.
- [42] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.
- [43] Sindy Löwe, Peter O’Connor, and Bastiaan S. Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019.
- [44] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, Proceedings of Machine Learning Research, 2020.
- [45] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2013.
- [46] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, 2017.
- [47] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 2019.
- [48] Yanlin Qian, Ke Chen, Jarno Nikkanen, Joni-Kristian Kamarainen, and Jiri Matas. Recurrent color constancy. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.
- [49] Nguyen Ho Man Rang, Dilip K. Prasad, and Michael S. Brown. Raw-to-raw: Mapping between image sensor color responses. In *CVPR*, pages 3398–3405. IEEE Computer Society, 2014.
- [50] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.
- [51] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

- [52] Ikuro Sato, Hiroki Nishimura, and Kensuke Yokoi. APAC: augmented pattern classification with neural networks. *CoRR*, 2015.
- [53] Lilong Shi and Brian Funt. Re-processed version of the gehler color constancy dataset of 568 images.
- [54] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illuminant estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, 2016.
- [55] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.
- [56] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 2014.
- [57] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. 2019.
- [58] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. abs/2005.10243, 2020.
- [59] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE Computer Society, 2011.
- [60] Joost van de Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE Trans. Image Process.*, pages 2207–2214, 2007.
- [61] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.
- [62] Tyler Vigen. Spurious correlations. Hachette books, 2015.
- [63] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.
- [64] Jin Xiao, Shuhang Gu, and Lei Zhang. Multi-domain learning for accurate and few-shot color constancy. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020.
- [65] Kai Y. Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *CoRR*, abs/2006.09994, 2020.
- [66] Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, and Guoping Qiu. End-to-end illuminant estimation based on deep metric learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3613–3622. IEEE, 2020.
- [67] Huanglin Yu, Ke Chen, Kaiqi Wang, Yanlin Qian, Zhaoxiang Zhang, and Kui Jia. Cascading convolutional color constancy. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 2020.
- [68] Xinyue Zhu, Yifan Liu, Zengchang Qin, and Jiahong Li. Data augmentation in emotion classification using generative adversarial networks. *CoRR*, 2017.