IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 18, 2022, accepted January 24, 2022, date of publication January 27, 2022, date of current version February 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3147252

Deep Spatio-Temporal Illuminant Estimation Under Time-Varying AC Lights

JUN-SANG YOO[1], KANG-KYU LEE[2], CHAN-HO LEE[2], JI-MIN SEO[3], AND JONG-OK KIM[2], (Member, IEEE)

[1]Computer Vision Laboratory, Samsung Advanced Institute of Technology, Suwon-si, Gyeonggi-do 16678, South Korea
[2]School of Electrical Engineering, Korea University, Seoul 02841, South Korea
[3]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Jong-Ok Kim (jokim@korea.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by the Korean Government through the Ministry of Science and ICT (MSIT) under Grant 2019R1A2C1005834; and in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2021-2018-0-01421.

ABSTRACT Artificial lights, which are powered by alternating current (AC), are ubiquitous nowadays. The intensity of these lights fluctuates dynamically depending on the AC power. In contrast to previous color constancy methods that exploited the spatial color information, we propose a novel deep learning-based color constancy method that exploits the temporal variations exhibited by AC-powered lights. Using a high-speed camera, we capture the intensity variations of AC lights. Then, we use these variations as an important cue for illuminant learning. We propose a network composed of spatial and temporal branches to train the model with both spatial and temporal features. The spatial branch learns the conventional spatial features from a single image, whereas the temporal branch learns the temporal features of AC-induced light intensity variations in a high-speed video. The proposed method calculates the temporal correlation between the high-speed frames to extract the effective temporal features. The calculations are done at a low computational cost and the output is fed into the temporal branch to help the model concentrate on illuminant-attentive regions. By learning both spatial and temporal features, the proposed method performs remarkably under a complex illuminant environment in a real world scenario in which color constancy is difficult to investigate. The experimental results demonstrate that the proposed method produces lower angular error than the previous state-of-the-art by 30%, and works exceptionally well under various illuminants, including complex ambient light environments.

INDEX TERMS Temporal color constancy, temporal correlation, AC light, high-speed video.

I. INTRODUCTION

Color constancy is a fundamental task in the fields of computer vision, computational photography and image processing [1], [2]. Its ultimate objective is to recover the intrinsic surface color by removing the effect of illuminant chromaticity [3]–[6]. In this regard, it is crucial to separate illuminant chromaticity from a digital image where surface and illuminant colors are mixed. Thus, it is a significant ill-posed inverse problem [7]–[10] that has been widely investigated.

Illuminant estimation has been primarily studied in the spatial domain [11]. Spatial pixels were aggregated and analyzed from a single image to infer the illuminant of a scene. The statistics-based approach exploits the assumption that the color distribution of image pixels is statistically achromatic [12]–[16]. Although this approach is widely used in commercial devices, it is not effective for narrow color distributions, such as regions with uniform color. However, the physics-based approach estimates illuminant chromaticity by applying the dichromatic model to a spatial image [17]–[21]. Unlike the statistics-based approach, the physics-based approach is effective for narrow color distributions. However, determining the accurate model parameters is challenging owing to the severely ill-posed nature of the problem addressed by the physics-based model. The performance of this model critically depends on the number of specular pixels that contain strong specular reflections [17], [19], [21], [22],

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

making it practically difficult to extract genuine specular pixels from an image [19], [21].

Recently, the dichromatic model-based illuminant estimation in the temporal domain has been studied under time-varying alternating current (AC) lights [23]–[25]. In these studies, experiments were performed using temporal samples having identical diffuse reflections instead of spatially similar pixels. These methods show superior performances owing to the per-pixel estimation performed using the dichromatic model.

As imaging technologies have advanced significantly in recent years, high-speed consumer digital cameras have become widely available at low prices. These days, they are even available on smartphones [26], [27]. The high-speed capture capability facilitates capturing quick changes in motion and illumination [28], [29]. For example, high-speed cameras can capture rapid time-varying illumination flickers (caused by AC), which are imperceptible to human beings.

Conventional color constancy studies have exploited spatial pixels and have shown marginal performances in public illuminant environments with a mixture of various AC illuminants. In this paper, we perform illuminant estimation from the temporal flickers of AC lights using deep learning to overcome the limitations of color constancy. We assume that the temporal intensity variations caused by AC lights might provide an important cue regarding illuminant colors, as has already been validated in [23]–[25].

Unlike in the conventional methods, we estimate the illuminant color using a deep neural network that fuses spatial and temporal features in an end-to-end manner. We aim to learn the high-speed temporal variations of pixel intensity deeply. The outline of the proposed method is shown in Fig. 1. The proposed network takes considers spatial and temporal information and is composed of two sub-networks, *i.e.*, temporal and spatial branches. In the dichromatic reflection model, the high-speed AC variations of temporal samples at a fixed location are closely related to the illuminant color [25], [29]. Therefore, high-speed video frames are fed into the temporal branch to extract temporal features from the illuminant AC variations. The non-local temporal correlations between the video frames are calculated in advance to ensure that the model learns the temporal features more efficiently. This allows the temporal network to concentrate on attentive temporal features. On the contrary, the spatial branch learns the spatial features from a single image.

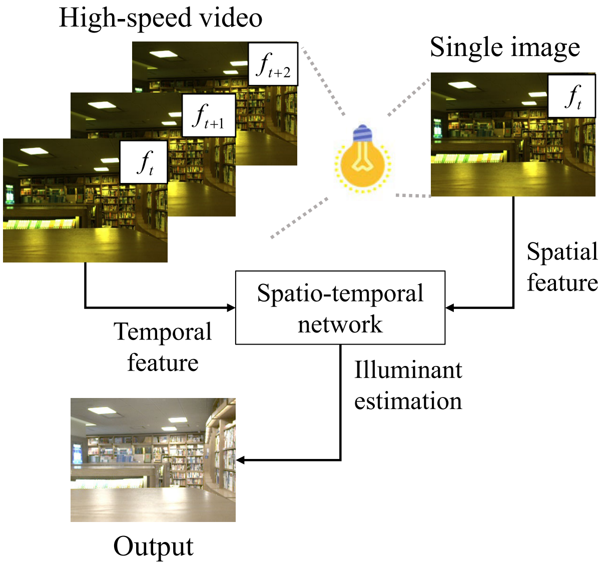


FIGURE 1. Outline of the proposed method. The proposed spatio-temporal network learns both temporal and spatial features to estimate the target illuminant using high-speed video as the input.

Diverse experiments performed using the dataset showed that the spatial features are useful for a simple illuminant environment in which only a single AC light exists, whereas the temporal features are advantageous in complex mixed-illuminant (or ambient) environments. To benefit from their respective advantages, the model learns the temporal and spatial features together, and combines them to estimate an illuminant. For performance evaluation, we construct a new high-speed AC dataset composed of scenes that can be classified into two groups: the first group is composed of scenes captured in a closed environment with one AC light, and the other consists of common indoor scenes under ambient light. The experimental results show that the proposed spatio-temporal network outperforms the previous state-of-the-art methods, and that it operates robustly under various illuminant environments.

The main contributions of this paper can be summarized as follows:

- The color of an AC illuminant is estimated by learning the temporal variations of illuminant intensity. This is realized via deep learning using video frame inputs.
- To the best of our knowledge, our approach to exploit temporal feature learning is the first trial of its kind in the field of color constancy. We propose a spatio-temporal network for illuminant estimation to make the model learn spatial and temporal features simultaneously. We observe that high-speed temporal features are useful for illuminant learning, and spatial features can further improve the performance of the proposed network.
- The non-local temporal correlation is calculated before feature extraction. This enables the temporal branch network to learn illuminant-sensitive regions easily, leading to efficient illuminant learning.
- To demonstrate the robustness of the proposed method, we constructed a new dataset composed of laboratory and real-world videos. Using this dataset, we demonstrate that the proposed method operates robustly under various illuminant environments.

The rest of the paper is organized as follows: In Section 2, we review related works found in literature on temporal and deep learning-based color constancy. In Section 3, we discusses the color constancy on temporal domain and the motivation of our work is presented. In section 4, we describe our proposed method in detail. Section 5 describes the dataset we used. Experimental results and ablation study are given in Section 6. Conclusion and future directions are given in Section 7.

II. RELATED WORKS

A. TEMPORAL COLOR CONSTANCY

Some color constancy methods [23]–[25], [30] that commonly rely on the dichromatic reflection model were proposed based in the temporal domain. Based on the assumption that the chromaticity of the incident light is constant between consecutive frames, the authors of [23] and [24] derived the chromaticity of the incident light using the dichromatic model. Furthermore, the authors of [25] specifically considered time-varying AC lights. The temporal illumination variations of the lights were captured using high-speed cameras and exploited for estimating dichromatic planes. The three aforementioned studies commonly applied temporal samples in a video sequence to the dichromatic reflection model. However, studies that estimate illuminants by training a model using video sequences are scarce. Thus, the method proposed in this study is a novel approach to use a spatio-temporal neural network. The conventional non-deep learning methods [23]–[25] are highly sensitive to temporal noise; in contrast, the proposed deep learning method performs better in a noise-robust manner, even when weak and noisy AC lights are employed in practical indoor scenes.

B. DEEP LEARNING-BASED COLOR CONSTANCY

Recently, several convolutional neural network (CNN) based color constancy studies have been conducted [31]–[38]. Oh and Kim [32], illuminant estimation was formulated as a classification problem that was solved using a CNN. The target illuminant was estimated by the weighted sum of probabilistic classification results using the network output, without using the illuminant estimation directly. Bianco *et al.* [39] estimated the illuminant locally using a CNN. The local estimates were refined using non-linear local aggregation, which yielded a global single estimate. Shi *et al.* [31] proposed a neural network consisting of two sub-networks, *i.e.*, hypotheses network (HypNet) and hypothesis selection network (SelNet). The former generated two hypotheses regarding the illuminant, and the latter chose the HypNet output that was closest to the groundtruth. Hu *et al.* [33] proposed an end-to-end illuminant estimation network whose outputs were pixel-wise illuminants and their confidences. A global illuminant was estimated by the weighted sum of individual pixel-wise illuminants using the confidence score. Bianco and Cusano [35] introduced unsupervised learning for color constancy. They used typical images available on the web instead of a conventional color constancy dataset and its illuminant labels. They assumed that the colors of these images were approximately balanced. Yu *et al.* [37] proposed a cascaded structure to improve the robustness of illuminant learning. A channel attention-based re-weighting module was adopted by Xiao *et al.* [38] to adapt to the camera-specific characteristics. Nevertheless, these previous methods only adopted image learning with regard to spatial aspects.

III. COLOR CONSTANCY IN TEMPORAL DOMAIN

In Asia and Europe, 60 Hz is used as standard frequency of AC electric power while 50 Hz is used in the Americas. For a standard frequency of 60 Hz, the AC electric power causes flickering at a rate of 120 times per second [28]. Although, various types of electric bulbs, such as incandescent, fluorescent, and compact fluorescent lamps (CFLs) have different fundamental principles for light emission, their brightness commonly flickers with a sinusoidal shape. Fig. 2 shows this AC variation. We can easily observe these AC flickers by plotting the intensity of the fixed point on high-speed frame axis. This can be an important cue for illuminants. In conventional dichromatic main stream studies like [19], [21], the authors used spatial pixels with identical diffuse chromaticities because the intensity of those pixels varies depending on the distance to a light source and this provides a cue for illuminant color. However, this spatial dichromatic line is very sensitive to noise, and it is highly challenging to extract spatial pixels with identical diffuse chromaticities.

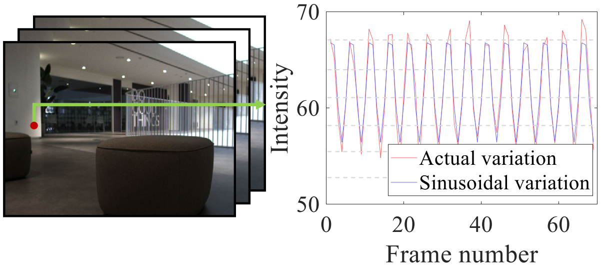


FIGURE 2. Temporal variation of a pixel in a high-speed video. Its variation is similar to that of a sinusoidal function with a frequency of 120 Hz. Note that this variation is captured at 150 fps.

The motivation for this study is that the temporal variations of an illuminant can be an important cue, and can be more advantageous than spatial variations. However, as discussed earlier, only a few studies have been conducted on temporal color constancy [23]–[25]. In this section, we show that the temporal variation of an individual pixel can be suitably exploited for color constancy using the dichromatic reflection model. The dichromatic reflection model can be expressed for the temporal domain as:

$$I_c(x,t) = m_d(x,t)\,\Lambda_c(x,t) + m_s(x,t)\,\Gamma_c(x,t), \quad c \in r,g,b \qquad (1)$$

where $I_c(x,t)$ is pixel intensity at a location x and time t, Λ_c and Γ_c are the diffuse and specular chromaticities, respectively, and, m_d and m_s are the diffuse and specular weights, respectively. In high-speed videos, captured scenes in adjacent high-speed frames are nearly identical. Thus, the pixel location x can be omitted by assuming a static video. Note that we capture the scenes at 150 frames per second (fps).

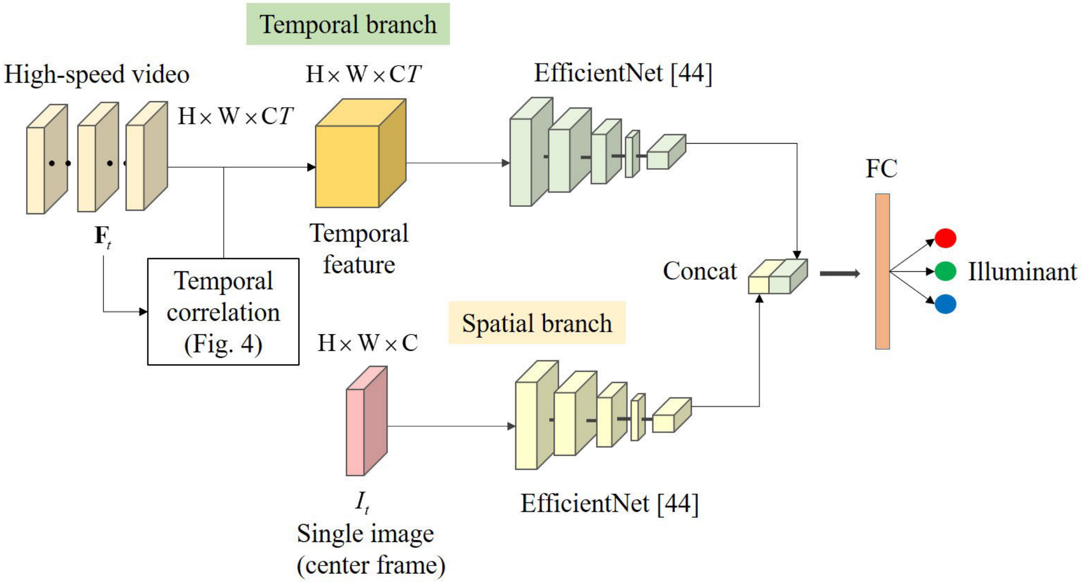


FIGURE 3. Proposed network structure composed of temporal and spatial branches.

Additionally, the diffuse and specular chromaticities are temporally constant. Thus, (1) can be simplified as follows:

$$I_c(t) = m_d(t)\,\Lambda_c + m_s(t)\,\Gamma_c \tag{2}$$

Then, the frame difference between $t + \Delta t$ and t is

$$I_c(t+\Delta t) - I_c(t) = \big(m_d(t+\Delta) - m_d(t)\big)\Lambda_c + \big(m_s(t+\Delta) - m_s(t)\big)\Gamma_c \tag{3}$$

Note that m_d reflects the intensity of the incident light. Yang *et al.* [23], [24] assumed that $m_d(t+\Delta) - m_d(t)$ was negligible. Thus, the direct determination of an illuminant was possible for a normal-speed camera. However, in the study of Yoo and Kim [25], this term was not negligible at a high-speed temporal scale. In their method, (3) physically indicated a dichromatic plane, as it was the weighted sum of two 3 × 1 RGB vectors. Thus, by calculating the intersection of several planes from different samples, the illuminant can be estimated as these planes share a specular chromaticity.

However, the study of Yoo *et al.* had some limitations: First, each extracted temporal AC sample should have a different diffuse chromaticity. If some samples share a diffuse chromaticity, the intersection of their planes will likely be biased, leading to a low accuracy in the illuminant estimation. Additionally, their study is not suitable for local shadow or ambient light because illuminant chromaticity cannot be shared in these cases.

Thus, deep learning is adopted for robust temporal color constancy to overcome these limitations. Important temporal features are extracted through a CNN and used for illuminant estimation. As described earlier, the temporal relationship between samples is important for illuminant estimation. Yoo *et al.* represented the temporal relationships using dichromatic planes and their intersections. Motivated by this, we introduce a temporal correlation method that measures the correlation of temporal samples at a fixed location and exploit it for illuminant learning. The conventional method employed in [25] requires a cautious selection of pixels for dichromatic plane estimation. In contrast, the proposed network automatically learns temporally attentive features considering all the temporal correlations in an end-to-end manner.

IV. SPATIO-TEMPORAL NETWORK ARCHITECTURE

Previous studies demonstrated that the illuminant color can be estimated from a set of temporal samples using the dichromatic reflection model [23]–[25]. In this paper, we adopt a data-driven deep learning approach instead of the theoretical dichromatic one. The proposed network aims to extract the temporal features of AC variations along with spatial features that are useful for illuminant estimation. In the dichromatic reflection model, it is assumed that the difference between adjacent video frames at a fixed location indicates the AC variations exhibited by an illuminant because diffuse chromaticity is maintained constant on the temporal axis. Based on this observation, we construct a spatio-temporal network comprising temporal and spatial sub-networks, as depicted in Fig. 3. The temporal branch network mainly learns the temporal features of a time-varying AC illuminant. The high-speed video facilitates the observation of the AC flicker between adjacent frames. This AC flicker is learned through the temporal branch, and its feature is extracted. Before feature extraction, the temporal correlations between video frames are calculated using a non-local neural network, and the input image is weighted using the temporal correlations, as illustrated in Fig. 4. Then, the non-local weighted input frames are fed into the temporal branch, leading to more efficient feature extraction. The learned temporal features

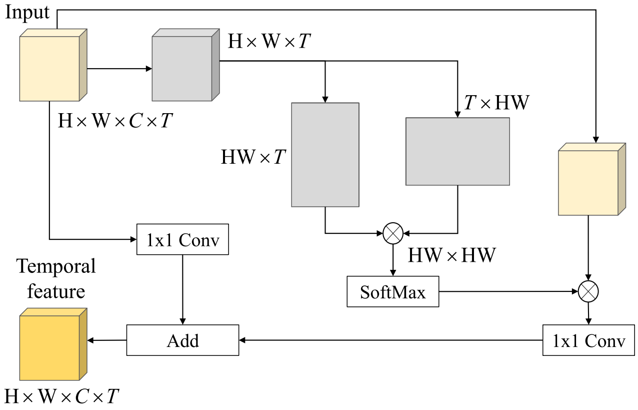


FIGURE 4. Calculation of the input temporal feature. Based on the non-local neural method [40], the temporal correlation weight is calculated and added to the input video feature.

concentrate on dynamically time-varying pixels, where the effect of illuminant flicker is more conspicuous. Note that this feature is similar to the AC pixels reported in [25]. The spatial branch of the model aims to learn spatial features from a single image as is accomplished in typical color constancy deep networks. We observed that this branch helps in improving the performance of the proposed method in simple illuminant environments such as a laboratory, whereas the temporal branch is advantageous for complex indoor environments. In the proposed method, these two branches work cooperatively to estimate the illuminant color, leading to a robust estimation of the target illuminant.

A. TEMPORAL CORRELATION WEIGHT

According to the dichromatic model, multiple temporal samples at a fixed location in a scene form a dichromatic plane (or line), and the dichromatic plane can be estimated for each pixel. The intersection of multiple dichromatic planes indicates an illuminant color. This indicates that the temporal relationship between non-local pixels plays a crucial role in illuminant estimation. In this paper, we propose a novel temporal correlation method that exploits the correlation of non-local features. It has been widely researched, and its effectiveness has been demonstrated [40]–[42]. Based on this approach, we estimate a temporal correlation feature. As shown in Fig. 4, the input high-speed video of size $H \times W \times C \times T$ is transformed to a gray-scale video whose size is $H \times W \times T$. Then, the 3D video signals are rearranged to the 2D matrix form of $HW \times T$. For correlation calculation, the 2D matrix is multiplied by its transposed version. The resulting correlation matrix whose size is $HW \times HW$ is normalized in a row-wise manner using the softmax algorithm [43], yielding the self-attention form; then, the input video frames are weighted using the correlations. Finally, both the weighted video frames and the original input frames undergo a 1×1 convolution, and the output features are combined. The resulting combination is fed into the temporal branch. The temporal features mainly capture temporally attentive regions such as the AC pixels in [25]. In other words, by reflecting the temporal correlation information in the input video frames, we can automatically obtain an illuminant map (to include AC variation information), which is very useful for illuminant estimation as demonstrated in [25]. The computation is simple because it is directly performed based on the input video frames and not the feature. Thus, it does not require $HW \times HW$ weight parameters unlike the model proposed in [40]. Furthermore, the computation is performed only once in the first layer.

B. NETWORK STRUCTURE

The proposed network is composed of two branches, as shown in Fig. 3. Given a current frame $\mathbf{I}_t$, the input video $\mathbf{F}_t$ can be defined using the surrounding frames as $\mathbf{F}_t = \{\mathbf{I}_{t-k}, \cdots, \mathbf{I}_t, \cdots, \mathbf{I}_{t+k}\}$. The size of $\mathbf{F}_t$ is $H \times W \times C \times T$, whereas that of $\mathbf{I}_t$ is $H \times W \times C$. Here, T is the number of frames in $\mathbf{F}_t$, which satisfies $T = 2k + 1$. $\mathbf{F}_t$ is used as an input for temporal learning, whereas $\mathbf{I}_t$ is used for spatial learning. In this study, EfficientNet-B0 [44] is used as the backbone for the spatial and temporal branches to extract deep illuminant features. The structure of EfficientNet-B0 is similar to that of MNasNet [45], as it uses a mobile inverted bottleneck MBConv block [46]. It is composed of six MBConv blocks and two additional convolution layers. Squeeze and excitation optimization [47] is added to it, and three model hyper-parameters (network depth, width, and resolution) are efficiently balanced using EfficientNet. This network demonstrates a powerful learning performance with a significantly low number of parameters. The temporally enhanced feature matrix is used as an input for the temporal branch, whereas the spatial branch uses $\mathbf{I}_t$. The output features of the final layer (the one before the fully connected layer) are concatenated, and fed into the fully connected layer. This fully connected layer produces an output vector with a size of 3×1. This vector is normalized to produce an estimated RGB illuminant vector, $\hat{\Gamma}$.

Given a ground truth illuminant, Γ_g, our proposed network is trained to minimize the following loss function:

$$\mathcal{L}\left(\hat{\Gamma}, \Gamma_g\right) = \arccos\left(\frac{\Gamma_g \cdot \hat{\Gamma}}{\left\|\Gamma_g\right\|\left\|\hat{\Gamma}\right\|}\right) \tag{4}$$

where $\|\cdot\|$ is L_2 norm, and $\mathcal{L}(\hat{\Gamma}, \Gamma_g)$ is the angular error between the estimated illuminant and the ground truth illuminant. In our proposed method, subnets have a common purpose, which is to extract deep illuminant features. The contribution of each branch depends on the illuminant characteristics of a scene. As shown in the following Ablation study, the spatial branch is more important for illuminant estimation in simple and monotonous illuminant environments, whereas the temporal branch plays a more important role for rather complex illuminant environments (mixed AC lights). This indicates that the proposed network adaptively learns illuminant features using the illuminant characteristics of a scene.

An input image can be corrected using the estimated illuminant. The proposed method applied the von Kries model [48], which scales the channel values of each pixel using the corresponding values of the normalized estimated illuminant.

V. DATASET

The proposed method exploits the high-speed video captured under AC light sources. The datasets commonly used in previous methods such as SFU Laboratory [53] and Gehler-Shi datasets [54] are composed of single shot images, thus they are not suitable for the proposed method. Recently, the INTEL-TAU dataset [55] has been published. It contains 7022 images in total, which is the largest number of images for a color constancy study. The dataset of [25] satisfies the requirements of the proposed method, but it was obtained in a closed environment using a matte cloth. Thus, the effect of external ambient light was not considered, and its scenes were monotonous and simple. To reflect the complex illuminant environments in the real world, the dataset of [25] was extended by adding the scenes of open spaces captured in the indoor environment. An additional dataset was constructed using a Sentech STC-MCS43U3V high-speed camera at 150 fps and an exposure time of 1/300, which is identical to the setting in [25]. As white balance is generally applied immediately after demosaicking in the image signal processing pipeline, the raw video is first linearized, and then demosaicked using the camera software provided. The demosaicked images are resized from 740 × 540 to 240 × 180 to reduce the computational cost.

The total number of scenes in the dataset is 225, *i.e.*, 150 scenes (66.7%) for training and 75 scenes for evaluation (33.3%). The set is categorized into two groups: one consisting of scenes captured in closed environments (33.3%), and the other consisting of scenes captured in public open indoor environments (66.7%). The number of frames in our video is 30, which is much more than the number of frames we use in the proposed method ($T = 5$). To the best of our knowledge, this is the first video-based dataset in color constancy studies that is captured under AC light sources.

With regard to the first group, the scenes are captured using a set-up identical to that used in [25], *i.e.*, the external ambient light is completely blocked using a matte cloth. Thus, it is a simple and ideal environment. Two types of AC lights, *i.e.*, incandescent and fluorescent lights, are used to produce equal numbers of scenes. The top row of Fig. 5 shows the sample scenes of the first group. The number of objects in a scene is at least three to guarantee the performance of the statistical methods. These scenes include various objects, such as metal, rubber, textile, stone, and plastic objects.

In contrast, the scenes in the second group are captured in indoor public places, such as a library, cafe, hotel and museum. These types of scenes are not included in the dataset of [25]. These scenes are mostly captured under ambient light, such as primary AC light with daylight or secondary AC lights. Although such scenes are challenging with regard to illuminant estimation, they are very common in practical scenarios. As shown at the bottom of Fig. 5, these scenes are affected by the mixture of various light sources, including the sun (daylight). We expect that the illuminant estimation performance can be evaluated more practically using this proposed dataset.

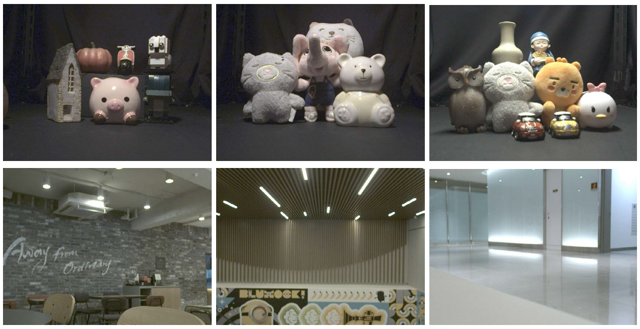


FIGURE 5. Some samples from the proposed dataset. To reflect the complex illuminant environments of the real world, the dataset is composed of various scenes under ambient light. The number of colors in a scene is more than four to guarantee desirable performance of the existing methods to a certain extent.

Note that all the scenes are captured additionally with a color checker for measuring the ground truth illuminant. The proposed method is evaluated and analyzed using this dataset. The details of these studies are described in the following section.

VI. EXPERIMENTAL RESULTS

The experiments are performed on a computer powered by an Intel i7-8700 CPU with 16GB RAM and two NVIDIA GTX 1080 Ti GPUs. The Adam optimizer is used for training the network at a learning rate of 0.001. We use five frames from a video input; thus, $T = 5$. We concatenate consecutive video frames as the network input of the temporal branch. The details of the experimental results are described in the following subsections.

A. COMPARISON WITH CONVENTIONAL METHODS

Table 1 compares the proposed method with conventional state-of-the-art methods. The performance is measured using the angular error given by $e = \mathcal{L}\left(\hat{\Gamma}, \Gamma_g\right)$. The conventional color constancy methods can be categorized into four groups: statistics-, gamut-, physics-, and learning-based method. Physics-based methods generally use the dichromatic reflection model. While IIC [19] and ICC [21] rely on a dichromatic line estimated from spatial samples on specular regions, Yoo and Kim [25] estimated a dichromatic plane for more accurate illuminant estimation by exploiting the AC pixels on the temporal axis. Note that Bianco *et al.* [39] and FC4 [33] are deep learning-based methods, whereas fast Fourier color constancy (FFCC) [52] learns filter parameters using UV histograms. For fair comparisons, the performances of the single-image-based methods are averaged for five frames (which is the same as the proposed method). Further,

TABLE 1. Angular error comparisons of various conventional methods.

Method		Mean	Median	Trimean	Best-25%	Worst-25%	Closed	Ambient
Statistics-based	Gray world	4.08	3.42	4.20	1.27	8.34	6.28	2.97
	Max-RGB [49]	13.22	13.62	12.20	5.92	20.65	14.85	12.40
	Shades of gray [12]	5.56	4.40	5.11	1.40	12.61	5.65	5.12
	1st order grey edge [13]	10.34	9.94	8.41	2.69	18.84	7.72	11.65
	2nd order grey edge [13]	12.39	12.12	9.18	3.65	22.06	9.59	13.79
	Grey pixels [14]	8.17	6.41	6.10	1.79	17.88	7.19	8.65
Gamut-based	Pixel gamut [50]	8.62	7.59	5.48	2.71	16.36	9.30	8.28
	1st order gradient gamut [51]	10.69	9.48	9.25	3.95	18.06	12.10	9.99
	2nd order gradient gamut [51]	10.54	10.36	8.93	2.24	19.54	12.73	9.44
Physics-based	IIC [19]	4.25	3.28	2.93	1.03	9.08	3.94	4.41
	ICC [21]	3.47	2.72	4.92	1.10	7.53	5.17	2.61
	Yoo *et al.* [25]	4.60	3.75	4.60	1.59	8.91	3.65	5.07
Learning-based	Bianco *et al.* [39]	1.79	1.12	1.20	0.36	4.42	1.44	1.97
	FFCC [52]	1.42	0.68	0.22	2.52	1.04	0.19	2.04
	FC4 [33]	2.26	2.05	2.00	0.76	4.17	2.30	2.25
	Proposed	1.00	0.43	0.40	0.26	2.57	0.56	1.22

the learning-based methods are trained in the same way as the proposed method.

Table 1 presents the performances of the color constancy methods using several statistical angular errors: mean, median, trimean, best-25%, and worst-25%. Furthermore, the results of the mean angular error are further classified as those obtained for closed environments and those for ambient environments to evaluate the robustness against practical illuminant environments. As listed in Table 1, the proposed method achieves the lowest angular error with regard to both the mean and median. Fig. 6 shows the training curve of the loss function for the proposed network. The proposed method effectively learns illuminant features with a small number of epochs.

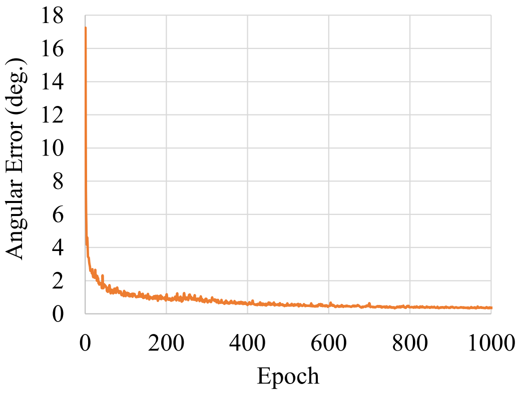


FIGURE 6. Training curve of the proposed loss function. As both the spatial and temporal branches share an identical angular loss function, the proposed network adaptively learns illuminant features with a small number of epochs.

B. ABLATION STUDY

To demonstrate the effectiveness of the proposed network, we conduct several experiments to determine the contributions of spatial and temporal information to the overall performance. As presented in Table 2, we index the ablation experiments from 1 to 5, where 5 corresponds to the proposed network. Ablation studies 1 and 2 are conducted to identify the significance of spatial and temporal information for illuminant estimation. Ablation study 1 uses the spatial branch of the network, whereas Ablation study 2 uses the temporal branch of the network. Ablation study 4 evaluates the effect of combining the spatial and temporal features. Finally, the effects of temporal correlation weight (Fig. 4) for video and image + video are evaluated in Ablation studies 3 and 5, respectively. Note that Wvideo is an abbreviation for the temporal-correlation-weighted video.

TABLE 2. Ablation study using various inputs.

	Total image		Closed	Ambient
Ablation study	Mean	Median	Mean	
1. Image	2.64	1.45	2.56	2.68
2. Video	1.24	0.45	0.61	1.56
3. Wvideo	1.15	0.37	0.90	1.27
4. Image + video	1.33	0.52	0.55	1.71
5. Image + Wvideo	1.00	0.38	0.56	1.22

1) BENEFIT OF HIGH-SPEED TEMPORAL FEATURES

To confirm the advantage of high-speed temporal features, we compare the performance obtained using a video with that obtained using a single image input by employing an identical network (EfficientNet-B0). Note that the only difference between the two networks for Ablation studies 1 and 2 is the channel number of the first layer. The network with video input employs the temporal branch without the temporal correlation module, and the model with a single image input uses the spatial branch shown in Fig. 3. The results of Ablation studies 1 and 2 (Table 2) show that learning using the high-speed video is considerably more effective than learning single images of closed and ambient environments. This suggests that the high-speed temporal variations exhibited by illuminants play a vital role in illuminant estimation.

2) EFFECT OF SPATIAL FEATURES

In Ablation study 1, we experimentally confirmed that high-speed AC flickers are an important cue for illuminant

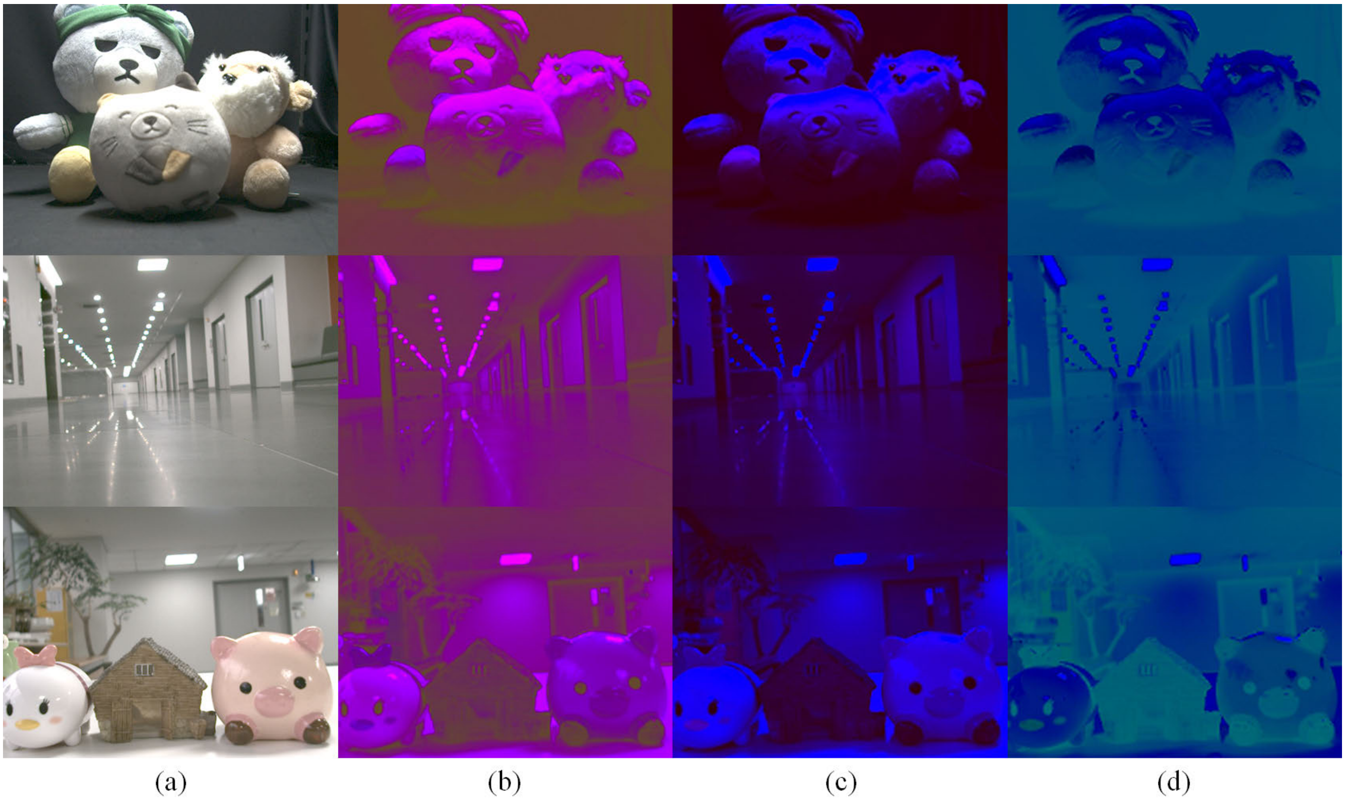


FIGURE 7. Visualization of the learned features (output of the first layer in the network) and comparison of the ablation studies in Table 2 (1 × 1 layer). (a) Original image, (b) spatial branch in Ablation study 1, (c) temporal branch in Ablation study 2, (d) temporal branch with correlation weight in Ablation study 5.

estimation. In Ablation study 4, we feed a single image and video to the network. For closed environments, the result demonstrates that the combination of 'video + image' is superior to the case where only the video is used, however, this combination is found to be inferior in the case of ambient environments. This indicates that the spatial feature is effective for a single uniform illuminant environment. However, it is not beneficial for complex illuminant environments. This can be also confirmed by comparing the results of Ablation study 3 with those of Ablation study 5. By adding spatial features to the temporal branch with correlations, the performance in closed environments is considerably improved. However, the performance gain is marginal in the case of ambient environments.

3) CONTRIBUTION OF TEMPORAL CORRELATION

In Section IV, we discussed the method for calculating temporal correlations. Adding a temporal correlation to the temporal branch makes it easy for the model to concentrate on attentive temporal features. By comparing the results of Ablation study 2 with those of Ablation study 3, we can observe that the performance achieved using the video as the input can be further improved by using temporal correlations. Similarly, upon comparing the results of Ablation studies 5 and 4, we can observe that the performance of the combined network is considerably better. The effect of temporal correlation is prominent for complex ambient environments, in which it is difficult to estimate illuminant chromaticity using spatial features only. The illuminant chromaticity can be accurately estimated by using the dynamic temporal fluctuations of AC bulbs and their correlations.

Fig. 7 compares the features of the first layer in the proposed network for various inputs, *i.e.*, (b) single image, (c) high-speed video, and (d) temporal-correlation-weighted high-speed video (Ablation studies 1, 2, and 3 in Table 2, respectively). It was observed that the learned features (b) and (c) are similar to the input image, (a). In contrast, the pixels with temporally specular-sensitive regions are well-captured for (d). The high-speed temporal feature corresponding to AC variations is exploited for extracting strongly illuminated regions in an image, and the consideration of the temporal correlation plays a key role in making the model concentrate on illuminant-attentive regions during the learning stage.

C. HIGH-SPEED VS. NORMAL VIDEO INPUT

Inspired by the minute variations of light from an AC light source in high-speed frames, we design the proposed network to learn temporal features for illuminant estimation. High-speed capture is essential in the proposed method. We believe that high-speed information can improve the optimal convergence of the deep network owing to the usefulness of

TABLE 3. Angular error comparisons using a normal video.

	Total image		Closed	Ambient
	Mean	Median	Mean	
Normal	1.42	0.92	1.04	1.60
High-speed	1.00	0.38	0.56	1.22

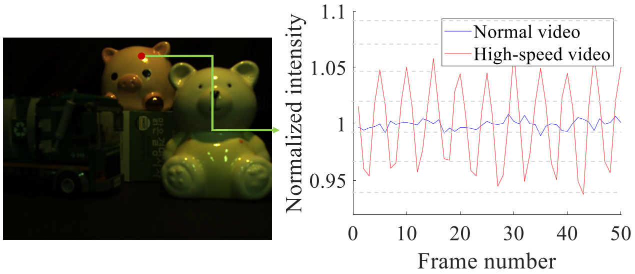


FIGURE 8. Pixel variations with normal (25 fps) and high-speed (150 fps) videos. By fixing the locations of camera and objects, we captured an identical region (denoted by the red point) in normal and high-speed videos.

the temporal features. Fig. 8 shows the variations of pixel intensity. The locations of the camera and objects are fixed to compare the variations of a pixel between normal and high-speed videos. Then, an identical pixel is captured at a normal frame rate (25 fps) and a high frame rate (150 fps). As shown in Fig. 8, the pixel corresponding to the high-speed case fluctuates sinusoidally, whereas the pixel corresponding to the normal case varies irregularly. This indicates that the variation of pixel intensity in the normal case is caused by noise, rather than the change in the brightness of an illuminant. The proposed network is trained using a standard video to demonstrate the effectiveness of high-speed AC variations. For a fair comparison, the standard video dataset is composed of scenes identical to those of the high-speed video dataset. Particularly, for the standard video experiment, the high-speed video is sub-sampled to a low frame rate. As presented in Table 3, it is observed that the proposed network with high-speed video input learns the temporal features better than the one with the standard video input. In other words, the high-speed sinusoidal variation is an extremely important cue for illuminant learning.

D. COMPLEX MOTION VIDEO

There would be complex motion scenes in real environments. To verify that the proposed method is robust to motion changes, 11 additional test scenes with moving objects were captured. Additional test scenes cover various moving situation, dynamic movements such as working and running, and small movements such as hand shaking. Table 4 compares the proposed method with the learning-based SOTA methods [33], [39], [52] using moving dataset. Compared with the result of static dataset, the performance of the proposed method is slightly inferior to FFCC. As the proposed temporal correlation concentrates on the varations of AC pixels, extreme motions can disturb estimating meaningful temporal correlation weight. However, the proposed network not only uses temporal correlation weight, but also uses temporal feature (high-speed video) and spatial feature (a single image). Furthermore, the frame rate of high-speed camera is 150fps, which is significantly faster than normal cameras. Therefore, even if the movement in the scene is complex, the difference between frames is relatively small compared to normal video, so its influence is also reduced. Therefore, the proposed method can achieve meaningful performance in the complex motion scene.

TABLE 4. Angular error comparisons in the complex motion scenes.

	Bianco *et al.* [39]	FFCC [52]	FC4 [33]	Proposed
Angular error	5.30	3.65	4.49	3.74

VII. CONCLUSION

We proposed a deep spatio-temporal network for performing illuminant estimation using a high-speed video as an input. Using this high-speed video, we could capture the minute temporal variations of AC light. We used these variations as an important cue for illuminant estimation. The proposed network comprises temporal and spatial branches for the efficient extraction of spatial and temporal features. We adopted temporal correlations and included them in the temporal branch to effectively concentrate on temporal features. As the existing dataset is insufficient to reflect real-world scenarios, we extended the conventional dataset by adding real-world scenes that were captured in complex illuminant environments. The experimental results showed that the proposed method performed robustly under various illuminant environments and that its performance was better compared with that of other state-of-the-art methods. In future work, the use of the deep spatio-temporal network can be extended not only for color constancy, but also for other tasks that can exploit the temporal characteristics of illuminants such as highlight removal.

ACKNOWLEDGMENT

(Jun-Sang Yoo and Kang-Kyu Lee contributed equally to this work.)

REFERENCES

[1] A. Byrne and D. R. Hilbert, *Readings on Color: The Science of Color*, vol. 2. Cambridge, MA, USA: MIT Press, 1997.

[2] H. R. Kang, *Computational Color Technology*. Bellingham, WA, USA: SPIE, 2006.

[3] L. T. Maloney and B. A. Wandell, "Color constancy: A method for recovering surface spectral reflectance," in *Readings in Computer Vision*. Amsterdam, The Netherlands: Elsevier, 1987, pp. 293–297.

[4] M. Ebner, *Color Constancy*, vol. 7. Hoboken, NJ, USA: Wiley, 2007.

[5] D. H. Foster, "Color constancy," *Vis. Res.*, vol. 51, no. 7, pp. 674–700, Apr. 2011.

[6] M. A. Hussain and A. S. Akbari, "Color constancy algorithm for mixed-illuminant scene images," *IEEE Access*, vol. 6, pp. 8964–8976, 2018.

[7] J. A. Worthey, "Limitations of color constancy," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 7, pp. 1014–1026, 1985.

[8] M. D. Fairchild and L. Peter, "Chromatic adaptation to natural and incandescent illuminants," *Vis. Res.*, vol. 32, no. 11, pp. 2077–2085, 1992.

[9] V. Agarwal, B. R. Abidi, A. Koschan, and M. A. Abidi, "An overview of color constancy algorithms," *J. Pattern Recognit. Res.*, vol. 1, no. 1, pp. 42–54, 2006.

[10] X. Huang, B. Li, S. Li, W. Li, W. Xiong, X. Yin, W. Hu, and H. Qin, "Multi-cue semi-supervised color constancy with limited training samples," *IEEE Trans. Image Process.*, vol. 29, pp. 7875–7888, 2020.

[11] A. Gijsenij, T. Gevers, and J. van de Weijer, "Computational color constancy: Survey and experiments," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2475–2489, Sep. 2011.

[12] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Proc. Color Imag. Conf.*, vol. 2004, no. 1. Springfield, VA, USA: SIST, Jan. 2004, pp. 37–41.

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

[14] K.-F. Yang, S.-B. Gao, and Y.-J. Li, "Efficient illuminant estimation for color constancy using grey pixels," in *Proc. CVPR*, Jun. 2015, pp. 2254–2263.

[15] M. A. Hussain and A. S. Akbari, "Color constancy adjustment using subblocks of the image," *IEEE Access*, vol. 6, pp. 46617–46629, 2018.

[16] M. A. Hussain, A. Sheikh-Akbari, and E. A. Halpin, "Color constancy for uniform and non-uniform illuminant using image texture," *IEEE Access*, vol. 7, pp. 72964–72978, 2019.

[17] G. D. Finlayson and G. Schaefer, "Solving for colour constancy using a constrained dichromatic reflection model," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 127–144, 2001.

[18] G. Schaefer, "Robust dichromatic colour constancy," in *Proc. Int. Conf. Image Anal. Recognit.* Berlin, Germany: Springer, 2004, pp. 257–264.

[19] R. T. Tan, K. Ikeuchi, and K. Nishino, "Color constancy through inverse-intensity chromaticity space," in *Digitally Archiving Cultural Objects*. Boston, MA, USA: Springer, 2008, pp. 323–351.

[20] J. Van De Weijer and S. Beigpour, "The dichromatic reflection model-future research directions and applications," in *Proc. VISAPP*, 2011, p. 11.

[21] S.-M. Woo, S.-H. Lee, J.-S. Yoo, and J.-O. Kim, "Improving color constancy in an ambient light environment using the phong reflection model," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1862–1877, Apr. 2018.

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

[23] Q. Yang, S. Wang, N. Ahuja, and R. Yang, "A uniform framework for estimating illumination chromaticity, correspondence, and specular reflection," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 53–63, Jan. 2011.

[24] V. Prinet, D. Lischinski, and M. Werman, "Illuminant chromaticity from image sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3320–3327.

[25] J.-S. Yoo and J.-O. Kim, "Dichromatic model based temporal color constancy for AC light sources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12329–12338.

[26] D. J. Griffiths, "Developmemt of high speed high dynamic range videography," Ph.D. dissertation, Dept. Mech. Eng., Virginia Tech, Blacksburg, VA, USA, 2017.

[27] J. Bonato, L. M. Gratton, P. Onorato, and S. Oss, "Using high speed smartphone cameras and video analysis techniques to teach mechanical wave physics," *Phys. Educ.*, vol. 52, no. 4, May 2017, Art. no. 045017.

[28] M. Vollmer and K.-P. Möllmann, "Flickering lamps," *Eur. J. Phys.*, vol. 36, no. 3, Apr. 2015, Art. no. 035027.

[29] M. Sheinin, Y. Y. Schechner, and K. N. Kutulakos, "Computational imaging on the electric grid," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6437–6446.

[30] N. Wang, B. Funt, C. Lang, and D. Xu, "Video-based illumination estimation," in *Proc. Int. Workshop Comput. Color Imag.* Berlin, Germany: Springer, Apr. 2011, pp. 188–198.

[31] W. Shi, C. C. Loy, and X. Tang, "Deep specialized network for illuminant estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2016, pp. 371–387.

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

[33] Y. Hu, B. Wang, and S. Lin, "FC4: Fully convolutional color constancy with confidence-weighted pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4085–4094.

[34] M. Afifi, B. Price, S. Cohen, and M. S. Brown, "When color constancy goes wrong: Correcting improperly white-balanced images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1535–1544.

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

[36] H.-H. Choi and B.-J. Yun, "Deep learning-based computational color constancy with convoluted mixture of deep experts (CMoDE) fusion technique," *IEEE Access*, vol. 8, pp. 188309–188320, 2020.

[37] H. Yu, K. Chen, K. Wang, Y. Qian, Z. Zhang, and K. Jia, "Cascading convolutional color constancy," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12725–12732.

[38] J. Xiao, S. Gu, and L. Zhang, "Multi-domain learning for accurate and few-shot color constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3258–3267.

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

[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2018, pp. 7794–7803.

[41] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 603–612.

[42] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Dec. 2019, pp. 3146–3154.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Dec. 2017, pp. 5998–6008.

[44] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[45] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. CVPR*, Jun. 2019, pp. 2820–2828.

[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2018, pp. 7132–7141.

[48] L. T. Sharpe and K. R. Gegenfurtner, *Color Vision: From Genes to Perception*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[49] E. H. Land, "The retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, 1977.

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

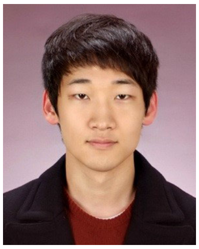
[51] A. Gijsenij, T. Gevers, and J. van de Weijer, "Generalized gamut mapping using image derivative structures for color constancy," *Int. J. Comput. Vis.*, vol. 86, no. 2, pp. 127–139, Jan. 2010.

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

[53] K. Barnard, L. Martin, B. Funt, and A. Coath, "A data set for color research," *Color Res. Appl.*, vol. 27, no. 3, pp. 147–151, Jun. 2002.

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

[55] F. Laakom, J. Raitoharju, J. Nikkanen, A. Iosifidis, and M. Gabbouj, "INTEL-TAU: A color constancy dataset," *IEEE Access*, vol. 9, pp. 39560–39567, 2021.



JUN-SANG YOO received the B.S. and Ph.D. degrees in electrical engineering from the School of Electrical Engineering, Korea University, Seoul, South Korea, in 2015 and 2021, respectively. He joined the Samsung Advanced Institute of Technology, Gyeonggi-do, South Korea, in 2021, where he is currently a Staff Researcher. His current research interests include sparse representations, super-resolution, and color constancy.

KANG-KYU LEE received the B.S. degree in electronic engineering from Korea University, Seoul, South Korea, where he is currently pursuing the master's and Ph.D. degree. His current research interests include intrinsic image decomposition, multi-exposure fusion, and color constancy.

CHAN-HO LEE received the B.S. degree from the Department of Electronic Engineering, Kwangwoon University, Seoul, South Korea, in 2019. He is currently pursuing the M.S. degree in electrical engineering with Korea University. His research interests include color constancy, highlight removal, noise reduction, and deep learning.

JI-MIN SEO received the B.S. degree from the School of Biomedical Engineering, Korea University, Seoul, South Korea, in 2018, and the M.S. degree in electrical engineering from Korea University, with a focus on signal processing and multimedia. He is currently pursuing the Ph.D. degree in electrical and computer engineering from Seoul National University. His current research interests include image/video processing, machine learning, and explainable artificial intelligence.

JONG-OK KIM (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Korea University, Seoul, South Korea, in 1994 and 2000, respectively, and the Ph.D. degree in information networking from Osaka University, Osaka, Japan, in 2006. From 1995 to 1998, he was an Officer at Korea Air Force. From 2000 to 2003, he was with the SK Telecom Research and Development Center and Mcubeworks Inc., South Korea, where he was involved in research and development on mobile multimedia systems. From 2006 to 2009, he was a Researcher with the Advanced Telecommunication Research Institute International (ATR), Kyoto, Japan. He joined Korea University, in 2009, where he is currently a Professor. His current research interests include image processing, computer vision, and intelligent media systems. He was a recipient of a Japanese Government Scholarship, from 2003 to 2006.