IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 22, 2021, accepted February 10, 2021, date of publication February 16, 2021, date of current version February 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059914

Learning-Based Illuminant Estimation Model With a Persistent Memory Residual Network (PMRN) Architecture

HO-HYOUNG CHOI[1], (Member, IEEE), AND BYOUNG-JU YUN[2]

[1]School of Dentistry, Advanced Dental Device Development Institute, Kyungpook National University, Daegu 41940, South Korea
[2]School of Electronics Engineering, College of IT Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: Byoung-Ju Yun (bjisyun@ee.knu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education under Grant NRF-2019R1I1A3A01061844, and in part by the NRF of Korea funded by the Ministry of Education under Grant NRF-2018R1D1A1B07040457.

ABSTRACT Since AlexNet, large deep convolutional neural networks (DCNNs) have been one of the major topics of interest in the field of computer vision, as well as bringing remarkable progress to the field. However, there has been little effort to use the DCNNs in realizing the mechanism of human memory. The human memory can be classified into three types: sensory memory, short-term memory and long-term memory. The short-term memory, also known as primary memory or active memory, is the information that humans are presently perceiving or thinking about, whereas the long-term memory refers to the persistent storage of information. In the mechanism of the human brain, the long-term memory enables the human vision to identify the actual color of an object effortlessly. In the computer vision, the DCNN-based illuminant estimation models are facing the long-term dependency problem as deeper networks encounter widening gaps between their earlier layers and later layers. Therefore, it is highly inspiring to apply the human long-term memory to the DCNN-based illuminant estimation models. The natural motivation of this article is to present a novel persistent memory residual network (PMRN) model which provides the DCNN with explicit access to persistent memory. The proposed PMRN architecture has two distinct units: a recursive unit and a gate unit. The two units combined serve to facilitate persistent memory access in a non-recursive fashion. The recursive unit has four residual blocks which are trained on the multiple-level image features on diverse receptive fields. The residual block outputs are concatenated and then fed into the gate unit. The proposed architecture keeps track of the recursive unit, deciding on how many of the earlier blocks to keep in reserve and how much of the image features to let the present block store. In this way, the proposed architecture contributes to solving the long-term dependency problem of conventional DCNNs. Comprehensive experiments support unparalleled performance of the proposed architecture in comparison to its counterparts and its potential to meet the needs of illumination estimation applications.

INDEX TERMS Human memory, long-term dependency problem, short-term memory, long-term memory, illumination estimation, persistent memory residual network (PMRN).

I. INTRODUCTION

The color constancy, known as one of the abilities of the human visual system (HVS), is a fundamental requirement for dealing with comprehensive computer vision issues. With color constancy, the computer vision is enabled to cope with negative illumination influence and perceive actual colors of an object effectively. Most computer vision-applied systems obtain diverse information from the actual colors of objects by performing color balancing of their input images. However, digital image devices have yet to equip their embedded photoelectronic sensors with an automatic function to correct illuminant colors. To address this problem, a lot of researchers have proposed state-of-the-art computational color constancy approaches by seeking to mimic the dynamic color balancing of the cones in the HVS [1]–[3]. These approaches are

The associate editor coordinating the review of this manuscript and approving it for publication was Liangtian Wan.

largely broken down into two categories: statistics-based and learning-based approaches.

The statistics-based methods make use of the predefined illuminant prediction models based on the grey-world hypothesis about the natural scene. The White-Patch [1] exemplifies this category in that the maximum response of three channels, red, green and blue is predefined as the achromatic color. Van De Weijer et al. [4] have tried to improve estimation accuracy by combining several statistics-based methods such as max-RGB [1], Gray-World (GW) [5], General Gray-World (GGW) [6] and Shades of Gray (SoG) [7]. In contrast, learning-based algorithms use network models which learn image patch data to perform illumination inference. Today, these algorithms are widespread in the computer vision community now that they radically outperform their statistic-based counterparts. Some of these learning-based methods have limitations when it comes to shallow network models which learn handcrafted image features [8]–[10]. However, recent learning-based approaches [11]–[14] have adopted the DCNNs which learn hierarchical feature maps. This work delves deeper into the DCNN-based approach and its implications. Bianco *et al.* [11] propose a learning-based method where non-overlapping image patches are provided for the CNN to estimate local illuminations, combine the results and estimate global illumination. Lou *et al.* [13] propose a network architecture which brings the global-level semantic context to the illumination estimation, and their method uses the whole image as input data. However, their architecture finds it hard to accurately estimate local semantic regions, which has vital influence on illuminant estimation results. Hu *et al.* [14] present a segment-wise approach, known as FC4, which utilizes the confidence maps of individual image patches to predict the color of the source illuminant. The FC4 model that is trained on semantic information delivers the smoothed results in the local regions, but finds it difficult to identify small objects. When the network mistakes small objects as noisy regions, the confidence maps are misled into masking the small objects, resulting in inaccurate local estimations. Afifi [10] introduces an algorithm with the semantic mask that proves effective in achieving satisfying computational color constancy. This approach requires preparing the algorithm to have the semantic mask in advance, which involves many complicated process steps. More recently, Choi *et al.* [15], [16] propose novel approaches by combining residual networks and dilated convolution to address the color constancy issue. These methods surpass their state-of-the-art counterparts, but come with a serious drawback: huge computation burden since the network is required to learn immense amounts of extra parameters using dilated convolution. With the learning-based approaches, the network models are made up of independent layers in sequence i.e. each layer affecting the next layer, which accounts for the short-term memory or the restricted long-term memory. Recently, it is published in neuroscience that the human brain preserves and stores in neocortical circuits what they acquire or are informed previously [17]. Unlike the human brain, the learning-based color constancy models do not have the mechanism to realize persistent memory.

Inspired by the unresolved challenges and opportunities for progress, this article presents a novel illuminant estimation model by adopting a large deep PMRN which has a memory block. The most interesting part of the proposed architecture is that a memory block consists of a recursive unit and a gate unit, which is intended to provide explicit access to persistent memory. The recursive unit has four residual blocks each of which learns the multiple level image features on diverse receptive fields in the mechanism of short-term memory. The residual block outputs are concatenated and then fed into the gate unit, which functions in a non-linear manner to implement persistent memory. In this way, the proposed PMRN realizes both short-term memory and long-term memory. In summary, the notable contributions of this article are as follows:

◊ Using the memory block in implementing the gating mechanism to navigate the long-term dependency challenge. Each residual block feeds its respective weights into the gate unit as its input and the gate unit is trained on those inputs. At the same time, the architecture keeps track of the recursive unit by deciding on how many of the previous blocks to keep in reserve and how much of the image features to let the present block store.

◊ With a very deep end-to-end PMRN for color constancy task, the proposed architecture produces mid- and high-frequency signals and facilitates the maximum possible network flow. To the best of our knowledge, this work is the first to bring the mechanism of human long-term memory to the illuminant estimation domain by building a novel PMRN.

◊ The proposed PMRN surpasses existing illumination estimation approaches by overcoming their limitation and estimating multiple illuminants, and accordingly demonstrating the latest performance. The proposed architecture is a single PMRN model and has robust learning competence.

II. RELATED WORK

A. ILLUMNATION ESTIMATION

A captured image is represented by the image formation model which has three factors: the light spectral distribution, e(λ), the surface reflectance at a specific location x,s(x, λ), and the camera spectral sensitivity function, c (λ) = $\{R(\lambda), G(\lambda), B(\lambda)\}^T$. In this model, λ refers to the wavelength of the light source. In obedience to the theory of Lambertian reflectance model, the integral calculus is used to approximate the pixel intensity value of the captured image, $\rho = [R, G, B]^T$, by calculating the light spectral distribution, the surface reflectance at a specific location and the camera spectral sensitivity function inside the human-visible spectrum [18], which goes as follows:

$$\rho_z(x) = m(x) \int_\omega e(\lambda)\, s(x,\lambda)\, c_z(\lambda)\, d\lambda; \quad z = \{R, G, B\} \tag{1}$$

where ω refers to the human-visible spectrum and $m(x)$ represents a diagonal matrix of the Lambertian reflectance shading of three channels, red, green and blue. Supposing that the perceived object color under one or multiple light sources is determined by the light spectral distribution, e(λ), and the camera spectral sensitivity function, c (λ), ignoring the influence of the imaging device, the light source e can be inferred and described as follows:

$$\mathrm{e} = \begin{pmatrix} e_R \\ e_G \\ e_B \end{pmatrix} = \int_\omega e(\lambda)\, s(x,\lambda)\, c_z(\lambda)\, d\lambda \tag{2}$$

In other words, the Lambertian reflectance model transforms the three factors of the image formation model into the pixel intensity value of the image, ρ, which in whole or in part implies and thus is used to infer the color of the light source, e, of the captured image.

B. BASIC MEMORY NETWORK

Let f_{ext} refer to the feature extraction operator and B_0 denote the extracted features. Basically, a memory network uses a convolutional layer to extract the features from input image data, x, and feed them into its first memory block (MB), which goes as follows [19]:

$$B_0 = f_{ext}(x) \tag{3}$$

Provided that the memory network consists of the stacked m memory blocks which process the feature maps, the feature map output, B_m, is a function of the m$^{\text{th}}$ memory block, M_m, described as follows:

$$B_m = M_m(B_{m-1}) = M_m(M_{m-1}(\ldots(M_1(B_0)))) + x \tag{4}$$

To learn the direct mapping, the memory network uses a convolutional layer in its ReconNet to rebuild the residual image [20]–[22]. Accordingly, the memory network can be formulated as follows:

$$\mathrm{y} = \mathrm{D}(\mathrm{x}) = f_{rec}(M_M(M_{M-1}(\ldots(M_1(f_{ext}(x)))))) + x \tag{5}$$

where f_{rec} represents the rebuild function and D refers to a function of the memory network. If a training set is defined as $\{x^{(i)}, \tilde{x}^{(i)}\}_{i=1}^{N}$ where N is the number of training image patches and $\tilde{x}^{(i)}$ is the ground truth of the input patch $x^{(i)}$, the memory network can compute a function of the train loss, expressed as follows:

$$\mathcal{L}(\Theta) = \frac{1}{2N}\sum\nolimits_{i=1}^{N}\left\|\tilde{x}^{(i)} - D(x^{(i)})\right\|^2, \tag{6}$$

where Θ refers to a set of the parameters.

III. THE PROPOSED PMRN APPROACH

In recent years, deep learning approaches have played a vital role in making huge advances in several vision tasks. This is possible mainly because of the progress in the capability to handle massive labeled training datasets. Still, the computer vision is facing absolute shortages of the labeled training datasets, among other image dataset problems, for illumination estimation, which thus calls for the color constancy innovation to estimate the most likely single or multiple illumination colors. Figure 1 illustrates the question (top) that illumination estimation research is seeking to answer (bottom) by publishing a lot of state-of-the-art learning-based network models and proving their models to outperform their statistics-based counterparts in terms of accuracy. Today, it is very pronounced in neuroscience that there are a lot of recursive connections in the neocortex of the human brain. Meanwhile, there has been little work to build the mechanism of recursive connections into the DCNN-based architecture, which has the potential to take illumination estimation to a higher level of accuracy. Inspired by the insights about the recursive connection mechanism in the human brain, this work presents a novel PMRN architecture which consists of a recursive unit and a gate unit. Figure 2 illustrates how persistent memory works in the proposed architecture. Every image feature from the recursive unit is concatenated and then fed into the gate unit.

This section dives deep into the proposed PMRN architecture. The proposed architecture has a memory block consisting of a recursive unit and a gate unit. The recursive unit operates in a non-linear manner, like human synapses which function in a recursive fashion [23], [24]. The recursive unit has four residual blocks (RB) that were originally proposed in ref. [25] and they fulfill robust learning, object recognition and image classification through the process of recursion. Let R refer to a function of the residual block, and H_m^{r-1} and H_m^r denote the input and the output of the r$^{\text{th}}$ residual block, respectively. Let F refer to the residual function (RF) with $\mathrm{r} = 1$ and $H_m^0 = B_{m-1}$, and W_m denote the set of weights on which the proposed architecture is trained. The residual block output is a function of the m$^{\text{th}}$ memory block, described as follows:

$$H_m^r = R_m\left(H_m^{r-1}\right) = F\left(H_m^{r-1}, W_m\right) + H_m^{r-1} \tag{7}$$

Every residual block has a skip connection, and three convolutional layers each of which has the pre-activation structure [26], which goes as follows:

$$\mathrm{F}\left(H_m^{r-1}, W_m\right) = W_m^3\tau\left(W_m^2\tau\left(W_m^1\tau\left(H_m^{r-1}\right)\right)\right); \quad W_m^i, i = 1, 2, 3, \ldots, \tag{8}$$

where τ is an activation function, ReLu [27], which is the following step after batch-normalization. Next, when the recursive unit generates multiple-level image features on receptive fields in a recursive fashion, the residual block takes and learns the image features, which illustrates how short-term memory works in the proposed architecture. Then, the short-term memories are ready for concatenation. Supposing that the recursive unit performs recursion R times, the r$^{\text{th}}$

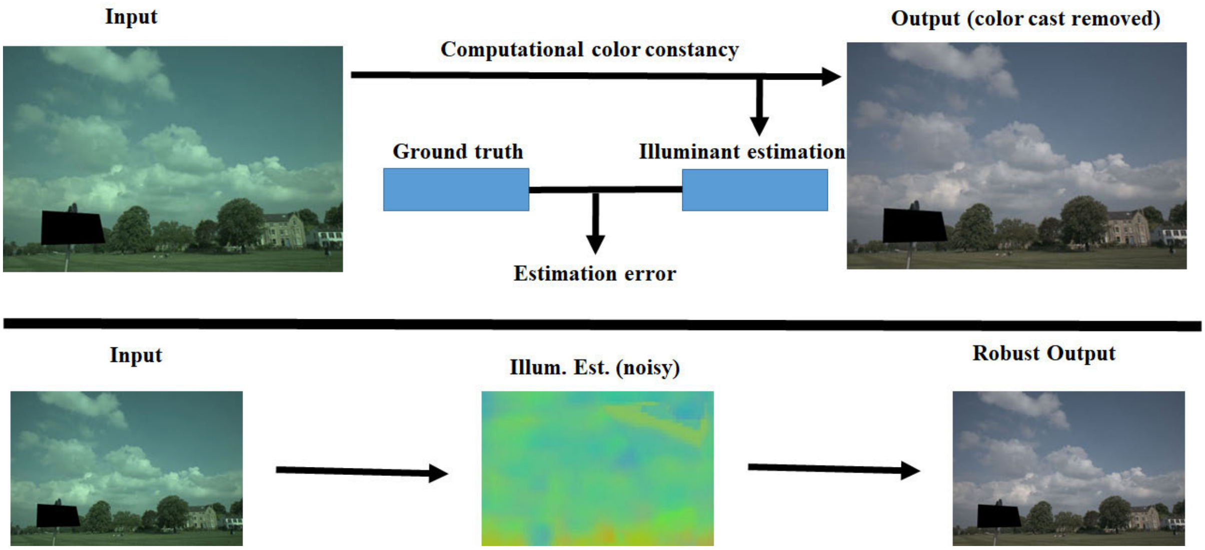


FIGURE 1. The convolutional neural network question (top) that learning-based approaches seek to answer (bottom).

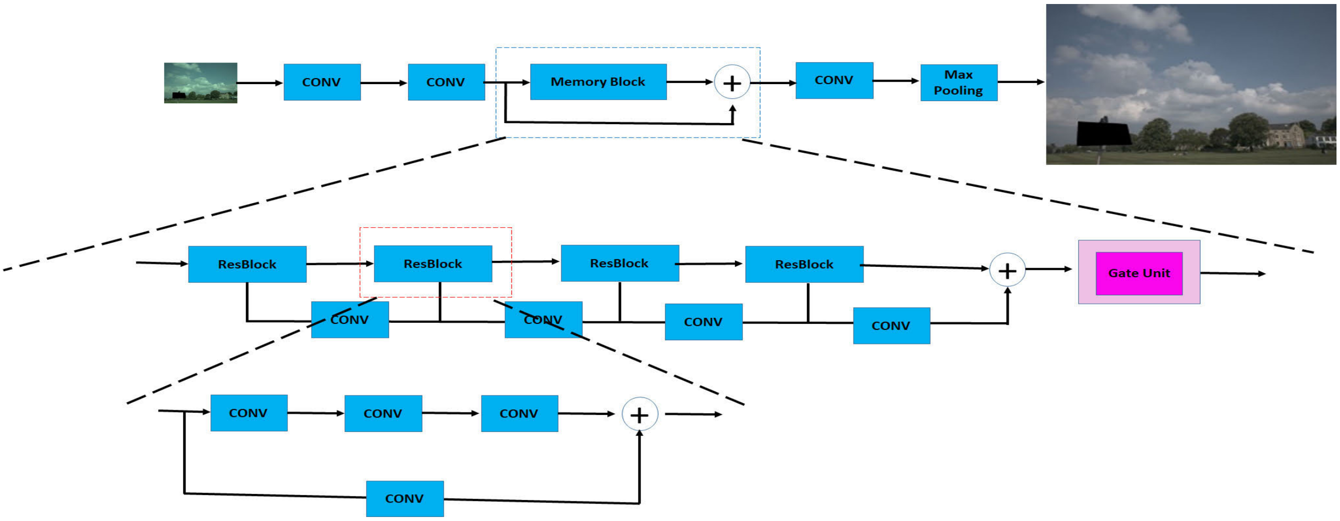


FIGURE 2. The structure of the proposed PMRN architecture.

recursion is expressed as follows:

$$H_m^r = R_m^r\left(B_{m-1}\right) = \underbrace{R_m(R_m(\ldots(R_m(R_{m-1}))\ldots))}_{r} \qquad (9)$$

where the $r - fold$ recursions are calculated and $\{H_m^r\}_{r=1}^R$ represents concatenation of multiple-level image features or short-term memories: $B_m^{short} = [H_m^1, H_m^2, \ldots, H_m^R]$. In addition, the proposed architecture learns what the (m-1)th residual block passes into the mth residual block at the present moment, which accounts for long-term memory, described as $B_m^{long} = [R_1, R_2, \ldots, R_{m-1}]$. Both short-term memory and long-term memory are concatenated and then fed into the gate unit as its input image features, which goes as follows:

$$B_m^{gate} = [B_m^{short}, B_m^{long}] \qquad (10)$$

The proposed architecture learns the weights of both short-term and long-term memory adaptively by means of the gating mechanism of the gate unit or a 1 × 1 convolutional layer. Supposing that f_m^{gate} is a function of the 1 × 1 convolutional layer parameterized by W_m^{gate} where B_m is the output of the mth memory block, the gating mechanism is modeled as follows:

$$B_m = f_m^{gate}\left(B_m^{gate}\right) = W_m^{gate}\tau(B_m^{gate}) \qquad (11)$$

Based on the weights of the long-term memory, the proposed architecture keeps track of the recursive unit by deciding on how many of the previous blocks to keep in reserve and how much of the image features to let the current block store. Therefore, the gating mechanism can be redefined as

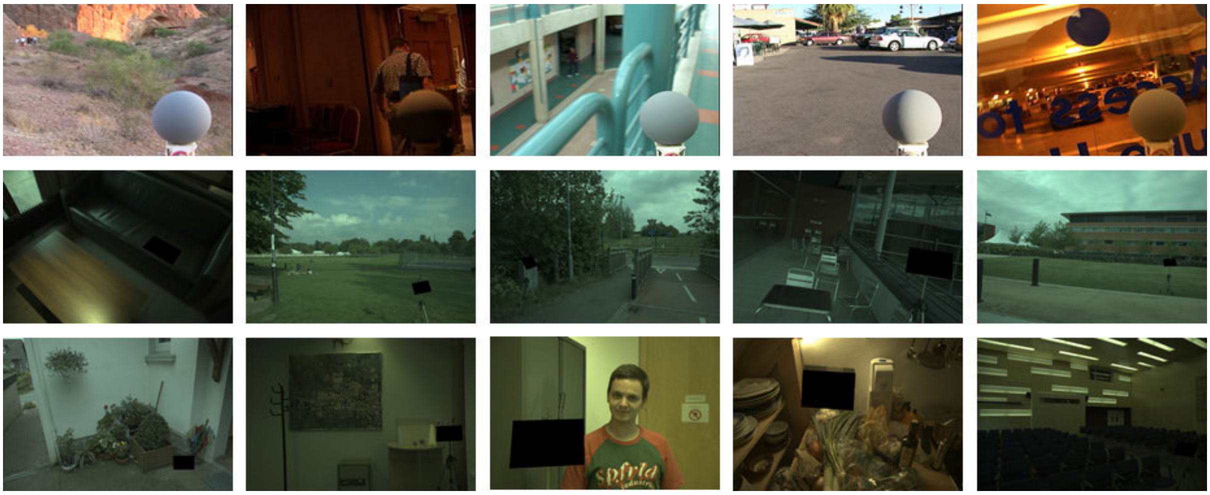


FIGURE 3. Sample images of each dataset: Gray-ball, NUS-8 camera and Shi's datasets.

follows:

$$B_m = M_m\left(B_{m-1}\right) = f_{gate}(\Big[R_m\left(B_{m-1}\right), \ldots, R_m^{(R)}\left(B_{m-1}\right), B_0, \ldots, B_{m-1}\Big] \qquad (12)$$

Based on the mathematical analyses above, the proposed PMRN is highly effective in taking illumination estimation to the next level of accuracy and efficiency by overcoming the long-term dependency problem and outperforming its conventional counterparts.

TABLE 1. Instances of different kernel sizes for learning time per epoch.

Instances	Filter size	Learning time (sec.)
Inst_1 ×1	1 ×1, 1 ×1, 1 ×1, 1 ×1	8.1
Inst_3 ×1	3 ×3, 1 ×1, 3 ×3, 3 ×3	9.5
Inst_3 ×3	3 ×3, 3 ×3, 3 ×3, 3 ×3	10.1
Inst_3 ×5	3 ×3, 5 ×5, 3 ×3, 3 ×3	10.8
Inst_5 ×5	5 ×5, 5 ×5, 5 ×5, 5 ×5	12.1

IV. EXPERIMENTAL RESULTS AND EVALUATIONS

This section discusses experimental datasets and their evaluations to verify the proposed architecture. The experiments use several standard datasets: Shi's dataset, NUS-8 camera dataset and Gray-ball dataset. The RAW images, originally taken with a higher-quality SLR camera without color adjustment in sRGB format [28], are reprocessed by Shi and Funt [29] into a 14-bit linear high-dynamic-range image format, not an 8-bit standard format. Another part of Shi's dataset includes 568 indoor and outdoor images, taken with Canon 5D and Canon 1D DSLR cameras. While taking the photos, the Macbeth Color Checker (MCC) chart is put up in every scene to ensure distinct contrast between the definite illumination color and the remainder on each resulting image, as well as facilitating accurate illumination estimation. The black level offset of the cameras is eliminated in advance, as suggested in ref. [29]. The NUS-8 camera dataset [30] is similar to Shi's dataset in that a digital SLR camera is used with the MCC chart in front, while taking photos of every scene. A difference is the numbers of cameras and images: total 8 cameras are used to take a massive number of images. Specifically, the dataset has 1,853 images with each camera taking approximately 200 images. The Gray-ball dataset [31] comprises approximately 11,000 images taken by a digital video camera with a neural gray ball in front, reflecting the color of the ambient light in every individual scene. The gray ball is in sight at all times. Figure 3 exemplifies the above three datasets. Starting with Shi's dataset, the proposed architecture resizes the images in RAW format into 512×512 input image patches and then crops the resized image patches into images in the max $(\mathrm{w}, \mathrm{h}) = 51,529$ pixels. The cropped image data is divided into three folds for three-fold cross validation. One fold is used to train the proposed architecture, another to validate, and the other to test. The proposed architecture takes the same process with the other datasets.

In parameter optimization experiments, this study selects two parameters: the initial learning rate and the kernel size, among several parameters, which are of vital importance for accuracy. The experiments are designed to determine the optimal initial learning rate for accuracy, while other parameters are fixed such as a weight decay of 5×10^{-5} and a momentum of 0.9. Figure 4 (a) compares average angular errors and Figure 4 (b) median angular errors at different initial learning rates. As a result, both the average angular error and the median angular error are minimized at the initial learning rate of 2.00E-4. The symbol "2.00E − 4" translates into 2×10^{-4}. In the proposed architecture, the convolutional layers take the input image data and extract the image features in the receptive field. The size of the receptive field depends on the size of the filter kernel at each convolutional layer. The narrow-sized filter kernel has advantages: a decreasing number of parameters and faster processing, but the disadvantage

TABLE 2. Comparison of the state-of-the-art learning-based and the proposed methods in terms of angular error with the use of Shi's dataset.

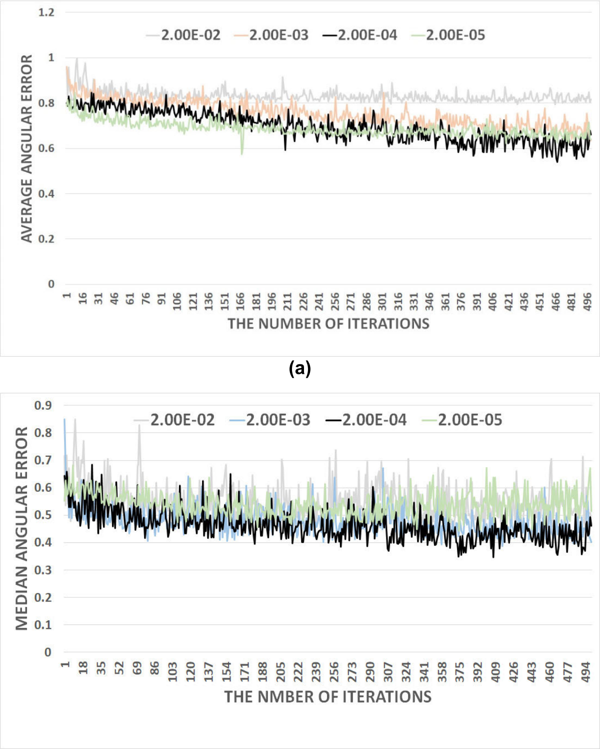
Methods	Mean	Median	Trimean	Best-25%	Worst-25%
SVR Regression Method[33]	8.08	6.73	7.19	3.35	14.89
Edge-based Gamut[34]	6.52	5.04	5.43	1.90	13.58
Bayesian Method[28]	4.82	3.46	3.88	1.26	10.46
Natural Image Statistics[35]	4.19	3.13	3.45	1.00	9.22
Intersection-based Gamut[33]	4.20	2.39	2.93	0.51	10.7
CART-based combination[36]	3.90	2.91	3.21	1.02	8.27
Spatio-spectral[37]	3.59	2.96	3.10	0.95	7.61
EM-based Method[38]	2.89	2.27	2.42	0.82	5.97
19-Edge Corrected-moment[39]	2.86	2.04	2.22	0.70	6.34
CNN based Method[40]	2.75	1.99	2.14	0.74	6.05
ED-based Method[9]	2.42	1.65	1.75	0.38	5.87
H. Zhan et al [41]	2.29	1.90	2.03	0.57	4.72
DS-Net[42]	2.24	1.46	1.68	0.48	6.08
SqueezeNet-FC4[14]	2.23	1.57	1.72	0.47	5.15
AlexNet-FC4[14]	2.12	1.53	1.64	0.48	4.78
Choi's method [15]	2.09	1.42	1.60	0.35	4.78
CMoDE[16]	2.05	1.06	1.42	0.29	4.50
Proposed method	**2.01**	**1.02**	**1.35**	**0.24**	**3.4**

TABLE 3. Comparison of the state-of-the-art learning-based and the proposed methods in terms of angular error with the use of Grey-ball dataset.

Method	Mean	Median	Trimean	Best-25%	Worst-25%
SVR-Regression Method	13.17	11.28	11.83	4.42	25.02
Bayesian Method	6.77	4.70	5.00	-	-
Natural Image Statistics	5.24	3.00	4.35	1.21	11.15
EM-based Method	4.42	3.48	3.77	1.01	9.36
CNN-based Method	4.80	3.70	-	-	-
Choi's method	4.03	1.88	2.60	0.61	10.77
CMoDE	3.28	1.70	2.27	0.53	6.69
Proposed method	**3.02**	**1.64**	**1.74**	**0.49**	**6.45**

is declining accuracy. So, the experiments in this work are designed to optimize the filter kernel size for each individual convolution layer. Table 1 compares learning times per epoch at different instances of filter kernel sizes and implies that a slight increase in the kernel size leads to a slight increase in the computational cost. As a result, the narrowest-size kernel filter, Inst_1 × 1, is the best performer in terms of efficiency. Figure 5 contrasts average angular errors and median angular errors at different instances of filter kernel sizes. The experiments use Shi's dataset to optimize the filter kernel size. Resultingly, both average angular error and median angular error are minimized at Inst_3 × 3 of the filter kernel size. The experiments use TensorFlow [32] with TITEN RTX D6 24G GPU support.

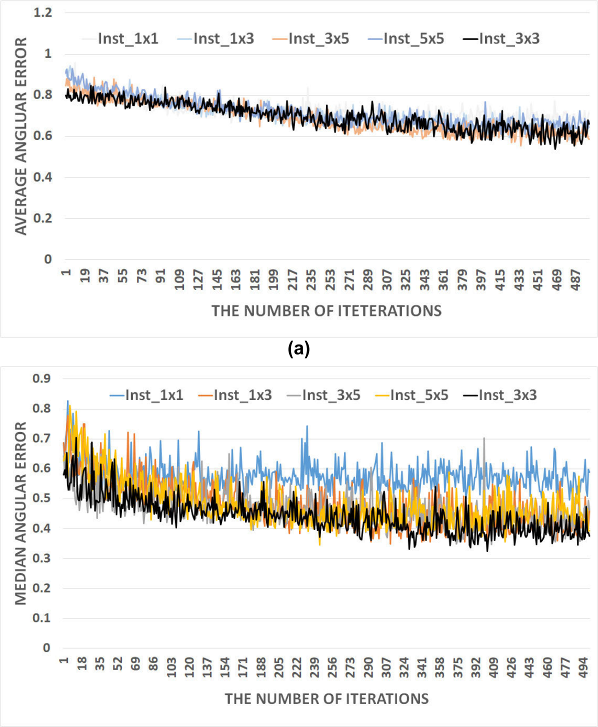
For peer comparison, the proposed method is compared with state-of-the-art learning-based methods using Shi's data. Table 2 is a comparison between state-of-the-art learning-based methods and the proposed method by angular error: average, median, trimean, best 25%, and worst 25%. Notably, Hu *et al.* [14] put forward a segment-wise approach, FC4, that uses the confidence maps of individual image patches in predicting the color of the source illuminant.



(a)

(b)

FIGURE 4. Comparison of (a) average angular error and (b) median angular error at different learning rates on the logarithmic space with the use of Shi's dataset.



(a)

(b)

FIGURE 5. Comparison of (a) average angular error and (b) median angular error at different filter kernel sizes on the logarithmic space with the use of Shi's dataset.

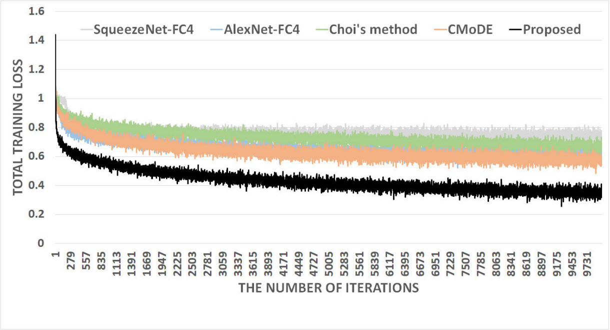


FIGURE 6. Comparison of the state-of-the-art and the proposed methods in terms of convergence of total training loss on the logarithmic space with the use of Shi's dataset.

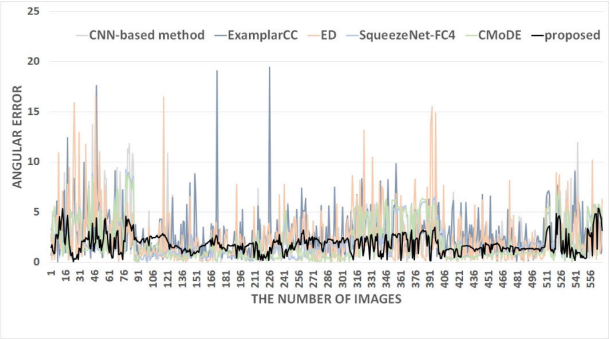


FIGURE 7. Comparison of select methods and the proposed method in terms of angular error distribution with the use of Shi's dataset.

In this approach, the network learns semantic information and builds the confidence maps. Resultingly, the network brings about the smoothed local regions on the images and causes some local estimations slightly unclear. The problem with this approach is that the network misleads the confidence maps into masking small objects which are not even noisy regions. With this approach, the network can hardly identify small objects. To address the accuracy problem with SqueezeNet-FC4 and Alex-FC4, Choi *et al.* [15], [16] come up with novel approaches, using the residual network and the dilated convolution. In result, these methods outperform their state-of-the-art counterparts in terms of accuracy, but come with a serious drawback: huge computation expense. To trigger the dilated convolution, these methods are forced to pad with zero to the left of the filter. This burdens the networks with extra massive parameters to learn. The approaches have another weakness: dependency on earlier layers in the networks. To resolve the problem and increase accuracy, several DCNN approaches have been developed as in ref. [43]–[45]. These approaches have been cited because of their high performance in the DCNN community. Nonetheless, there has been little work to use the DCNNs in realizing the mechanism of human memory. In this respect, this work proposes the PMRN approach by bringing the human memory mechanism and gating mechanism to illumination estimation. The memory block fulfils the gating mechanism, which serves to address the long-term dependency challenge facing previous CNN methods. In the memory block, a recursive unit is trained on multiple level image features, which represents the mechanism of short-term memory. The residual block outputs are

TABLE 4. Comparison of the state-of-the-art learning-based and the proposed methods in terms of angular error with the use of NUS-8 Camera dataset.

Method	PG	BF	SS	NIS	CM	CMo	PM
Camera	**Mean-Angular Error**						
Canon1Ds	6.13	3.58	3.21	4.18	3.18	3.05	**2.93**
Canon600D	14.51	3.29	2.67	3.43	2.35	2.21	**2.12**
FujiXM1	8.59	3.98	2.99	4.05	3.10	2.95	**2.76**
NikonD5200	10.14	3.97	3.15	4.10	2.35	2.23	**2.21**
OlympEPL6	6.52	3.75	2.86	3.22	2.47	2.32	**2.26**
LumixGX1	6.00	3.41	2.85	3.70	2.46	2.34	**2.25**
SamNX2000	7.74	3.98	2.94	3.66	2.32	2.18	**2.13**
SonyA57	5.27	3.50	3.06	3.45	2.33	2.21	**2.12**
Camera	**Median-Angular Error**						
Canon1Ds	4.30	2.80	2.67	3.04	2.71	2.43	**1.57**
Canon600D	14.83	2.35	2.03	2.46	2.19	2.05	**1.27**
FujiXM1	8.87	3.20	2.45	2.96	2.82	2.55	**2.01**
NikonD5200	10.32	3.10	2.26	2.40	1.92	1.66	**1.35**
OlympEPL6	4.39	2.81	2.24	2.17	2.12	1.81	**1.78**
LumixGX1	4.74	2.41	2.22	2.28	1.42	1.25	**1.22**
SamNX2000	7.91	3.00	2.29	2.77	1.32	1.15	**1.12**
SonyA57	4.26	2.36	2.58	2.88	1.65	1.53	**1.50**
Camera	**Trimean-Angular Error**						
Canon1Ds	4.81	2.97	2.79	3.30	2.69	2.52	**2.04**
Canon600D	14.78	2.40	2.18	2.72	2.33	2.21	**1.49**
FujiXM1	8.64	3.33	2.55	3.06	2.88	2.74	**2.12**
NikonD5200	10.25	3.36	2.49	2.77	1.95	1.90	**1.63**
OlympEPL6	4.79	3.00	2.28	2.42	2.18	1.95	**1.90**
LumixGX1	4.98	2.58	2.37	2.67	1.81	1.64	**1.62**
SamNX2000	7.70	3.27	2.44	2.94	1.65	1.51	**1.48**
SonyA57	4.45	2.57	2.74	2.95	1.91	1.75	**1.74**
Camera	**Mean of Best-25%**						
Canon1Ds	1.05	0.76	0.88	0.78	0.65	0.52	**0.49**
Canon600D	9.98	0.69	0.68	0.78	0.73	0.59	**0.47**
FujiXM1	3.44	0.93	0.81	0.86	0.75	0.81	**0.72**
NikonD5200	4.35	0.92	0.86	0.74	0.57	0.43	**0.42**
OlympEPL6	1.42	0.91	0.78	0.76	0.80	0.75	**0.73**
LumixGX1	2.06	0.68	0.82	0.79	0.65	0.53	**0.51**
SamNX2000	2.65	0.93	0.75	0.75	0.53	0.39	**0.33**
SonyA57	1.28	0.78	0.87	0.83	0.57	0.42	**0.4**
Camera	**Mean of Worst-25%**						
Canon1Ds	14.16	7.95	6.43	9.51	6.67	6.53	**6.35**
Canon600D	18.45	7.93	5.77	5.76	5.29	5.19	**5.15**
FujiXM1	13.4	8.82	5.99	9.37	5.64	5.53	**5.4**
NikonD5200	15.93	8.18	6.90	10.01	4.86	4.72	**4.66**
OlympEPL6	15.42	8.19	6.14	7.46	4.62	4.49	**4.4**
LumixGX1	12.19	8.00	5.90	8.74	5.74	5.55	**5.21**
SamNX2000	13.01	8.62	6.22	8.16	5.55	5.39	**4.58**
SonyA57	11.16	8.02	6.17	7.18	5.12	4.95	**4.60**

concatenated and then fed into the gate unit, which portrays the mechanism of long-term memory. Both short-term memory and long-term memory are concatenated and then fed into the gate unit in a non-linear function to realize persistent memory. As a result, the proposed architecture tops the the-state-of-the-art methods in the field of the illuminant estimation as in Table 2. Figure 6 contrasts the state-of-the-art methods and the proposed method in terms of total training losses and their respective learning behavior or convergence trends with the use of Kngma and Adam [46].

Resultingly, the proposed architecture tops the other methods by converging toward the lowest total training loss. Figure 7 compares angular error distributions of several high performers from Table 2: CNN, ExampleCC, ED, SqueezeNet-FC4, CMoDE, and the proposed architecture. In result, the proposed architecture tends toward the lowest angular error. To verify illumination invariant, another experiment is conducted to contrast conventional learning-based methods and the proposed architecture, with the use of Grey-ball dataset in terms of mean, median, trimean, best-25% and worst-25%. Table 3 summarizes the experimental results where the proposed PMRN architecture records the lowest angular error among its conventional counterparts. To verify camera invariant, another experiment is conducted to compare state-of-the-art conventional methods and the proposed PMRN architecture, with the use of the newest and widespread NUS-8 camera dataset [30] in terms of mean, median, trimean, best-25% and worst-25%. Table 4 summarizes the experimental results where the proposed architecture proves to outperform its latest counterparts regardless of the camera sensitivity.

V. CONCLUSION

This article presents a deep end-to-end persistent memory architecture with a memory block by bringing the human memory mechanism and gating mechanism into illumination estimation approaches. The memory block fulfils the gating mechanism to address the long-term dependency challenge with previous CNN methods. In the memory block, a recursive unit is trained on multiple level image features, which illustrates the mechanism of short-term memory. The residual block outputs are concatenated and then fed into the gate unit, which portrays the mechanism of long-term memory. Both short-term memory and long-term memory are concatenated and then fed into the gate unit in a non-linear function to realize persistent memory. Peer comparison supports the unparalleled superiority of the proposed PMRN architecture over its latest conventional counterparts. In the comparative experiments, various color constancy datasets are used such as Shi's dataset, Grey ball dataset and NUS-8 camera dataset. Still, more needs to be done to advance towards optimizing the CNN structure.

REFERENCES

[1] D. H. Brainard and B. A. Wandell, "Analysis of the retinex theory of color vision," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 3, no. 10, p. 1651, Oct. 1986.

[2] E. Y. Lam, "Combining gray world and retinex theory for automatic white balance in digital photography," in *Proc. 9th Int. Symp. Consum. Electron. (ISCE)*, Jun. 2005, pp. 134–139.

[3] X.-S. Zhang, S.-B. Gao, R.-X. Li, X.-Y. Du, C.-Y. Li, and Y.-J. Li, "A retinal mechanism inspired color constancy model," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1219–1232, Mar. 2016.

[4] 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.

[5] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Inst.*, vol. 310, no. 1, pp. 1–26, Jul. 1980.

[6] K. Barnard, L. Martin, A. Coath, and B. Funt, "A comparison of computational color constancy algorithms—Part II: Experiments with image data," *IEEE Trans. Image Process.*, vol. 11, no. 9, pp. 972–983, Nov. 2002.

[7] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Proc. IST/SID Color Imag. Conf.*, vol. 1, Jan. 2004, pp. 37–41.

[8] G. D. Finlayson, S. D. Hordley, and I. Tastl, "Gamut constrained illuminant estimation," *Int. J. Comput. Vis.*, vol. 67, no. 1, pp. 93–109, Apr. 2006.

[9] D. Cheng, B. Price, S. Cohen, and M. S. Brown, "Effective learning-based illuminant estimation using simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1000–1008.

[10] M. Afifi, "Semantic white balance: Semantic color constancy using convolutional neural network," 2018, *arXiv:1802.00153*. [Online]. Available: http://arxiv.org/abs/1802.00153

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

[12] J. T. Barron, "Convolutional color constancy," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 379–387.

[13] Z. Lou, T. Gevers, N. Hu, M. P. Lucassen, "Color constancy by deep learning," in *Proc. BMVC*, 2015, pp. 1–76.

[14] Y. Hu, B. Wang, and S. Lin, "FC^4: Fully convolutional color constancy with confidence-weighted pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4085–4094.

[15] H. H. Choi, H. S. Kang, and B. J. Yun, "CNN-based illumination estimation with semantic information," *Appl. Sci.*, vol. 10, no. 14, pp. 1–17, Jul. 2020.

[16] 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.

[17] J. Cichon and W.-B. Gan, "Branch-specific dendritic Ca2+ spikes cause persistent synaptic plasticity," *Nature*, vol. 520, no. 7546, pp. 180–185, Apr. 2015.

[18] K. Barnard, *Practical Colour Constancy*. Burnaby, BC, Canada: Simon Fraser Univ., 1999.

[19] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," 2017, *arXiv:1708.02209*. [Online]. Available: http://arxiv.org/abs/1708.02209

[20] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[22] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[23] P. Dayan and L. F. Abbott, *Theoretical Neuroscience*. Cambridge, MA, USA: MIT Press, 2001.

[24] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3367–3375.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, 2016, pp. 630–645.

[27] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 1–8.

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

[29] L. Shi and B. V. Funt. *Re-Processed Version of the Gehler Color Constancy Database of 568 Images*. Accessed: Oct. 26, 2020. [Online]. Available: http://www.cs.sfu.ca/~colour/data/shi_gehler

[30] D. Cheng, D. K. Prasad, and M. S. Brown, "Illuminant estimation for color constancy: Why spatial-domain methods work and the role of the color distribution," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 31, no. 5, pp. 1049–1058, 2014.

[31] F. Ciurea and B. Funt, "A large image database for color constancy research," in *Proc. 11th Color Image. Conf. Final Program*, 2003, pp. 160–164.

[32] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[33] W. Xiong and B. Funt, "Estimating illumination chromaticity via support vector regression," *J. Imag. Sci. Technol.*, vol. 50, no. 4, pp. 341–348, Jul. 2006.

[34] 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, nos. 2–3, pp. 127–139, Jan. 2010.

[35] A. Gijsenij and T. Gevers, "Color constancy using natural image statistics and scene semantics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 687–698, Apr. 2011.

[36] S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Automatic color constancy algorithm selection and combination," *Pattern Recognit.*, vol. 43, no. 3, pp. 695–705, Mar. 2010.

[37] A. Chakrabarti, K. Hirakawa, and T. Zickler, "Color constancy with spatio-spectral statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1509–1519, Aug. 2012.

[38] H. R. V. Joze and M. S. Drew, "Exemplar-based color constancy and multiple illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 860–873, May 2014.

[39] G. D. Finlayson, "Corrected-moment illuminant estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1904–1911.

[40] S. Bianco, C. Cusano, and R. Schettini, "Single and multiple illuminant estimation using convolutional neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4347–4362, Sep. 2017.

[41] H. Zhan, S. Shi, and Y. Huo, "Computational colour constancy based on convolutional neural networks with a cross- level architecture," *IET Image Process.*, vol. 13, no. 8, pp. 1304–1313, Feb. 2019.

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

[43] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, "Residual networks of residual networks: Multilevel residual networks," 2016, *arXiv:1608.02908*. [Online]. Available: http://arxiv.org/abs/1608.02908

[44] K. Zhang, Y. Su, X. Guo, L. Qi, and Z. Zhao, "MU-GAN: Facial attribute editing based on multi-attention mechanism," 2020, *arXiv:2009.04177*. [Online]. Available: http://arxiv.org/abs/2009.04177

[45] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*. [Online]. Available: http://arxiv.org/abs/1608.06993

[46] D. Kngma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

HO-HYOUNG CHOI (Member, IEEE) received the Ph.D. degree in mobile communication engineering from Kyungpook National University, Daegu, South Korea, in 2012. From 2014 to 2019, he was with Chungbuk National University as a Postdoctoral Researcher and a Contract Professor. He is currently a Visiting Professor with the School of Dentistry, Advanced Dental Device Development Institute, Kyungpook National University. His research interests include image processing, computer vision, machine vision, color constancy, tone mapping for HDR image, machine learning, and convolutional neural networks.

BYOUNG-JU YUN received the Ph.D. degree in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2002. From 1996 to 2003, he was with SK Hynix Semiconductor Inc., where he was a Senior Engineer. From 2003 to 2005, he was with the Center for Next Generation Information Technology, Kyungpook National University, where he was an Assistant Professor. Since 2005, he has been with the School of Electronics Engineering, Kyungpook National University, where he is currently an Invited Professor. His research interests include image processing, color consistency, multimedia communication system, HDR color image enhancement, biomedical image processing, and HCI.