Channel Attention GAN Trained With Enhanced Dataset for Single-Image Shadow Removal

RYO ABIKO AND MASAAKI IKEHARA, (Senior Member, IEEE)
Department of Electronics and Electrical Engineering, Keio University, Yokohama-shi, Kanagawa 223-8522, Japan
Corresponding author: Ryo Abiko (abiko@tkhm.elec.keio.ac.jp)

ABSTRACT Even today, where many deep-learning-based methods have been published, single-image shadow removal is a challenging task to achieve high accuracy. This is because the shadow changes depending on various conditions such as the target material or the light source, and it is difficult to estimate all the physical parameters. In this paper, we propose a new single-image shadow removal method (Channel Attention GAN: CANet) using two networks for detecting shadows and removing shadows. Intensity change in shadowed regions has different characteristics depending on the wavelength of light. In addition, the image acquisition system of the camera acquires an image in a state where the RGB values influence each other. Therefore, our method focused on the physical properties of shadows and the camera’s image acquisition system. The proposed network has a structure considering the relationship between color channels. When training this network, we modified the color and added some artifacts to the training images in order to make the training dataset more complex. These image processing are based on the shadow model, considering the camera image acquisition system. With these new proposals, our method can remove shadows in all ISTD, ISTD+, SRD, and SRD+ datasets with higher accuracy than the state-of-the-art methods. The code is available on GitHub: https://github.com/ryo-abiko/CANet.

INDEX TERMS Image restoration, deep learning, shadow removal, generative adversarial networks.

I. INTRODUCTION
When the light is blocked by some objects, the amount of light reflected from the target is reduced, and this becomes a shadow. It is very difficult to estimate the shadow-free image from the shadowed image since the property of the shadow changes by various conditions such as the material of the object, type of the light source, distance from the light source [1]. Images containing shadows degrade the accuracy of application methods such as face recognition, segmentation, and tracking [2]–[5]. Thus, removing shadows from images is an important task in computer vision.

In the early methods, shadow removal was based on a physical shadow generation model [8]–[11]. Those approaches require estimating various parameters such as reflectance for different materials and colors, light intensities for different locations, and direct and indirect light intensities. It is difficult to estimate these parameters accurately from a single image. Moreover, it cannot be said that the shadow models used in the papers are the perfect model of the actual shadow.

Therefore, even if the parameters can be estimated accurately, the performance of shadow removal based on the shadow model was limited.

On the other hand, the shadow removal method based on deep neural networks has recently produced good results [6], [7], [12]–[17]. There are two types of methods using deep neural networks, one is trained with a pair of shadow images and ground truth images [6], [7], [12]–[14], [16], [17], and the other is trained with a single image as a feature transformation task [15]. Although MaskShadowGAN [15] is a typical method that uses unpaired images for training, this method has a problem that images become blurred due to feature transformation, which has been improved by using GAN, but there is room for improvement. For methods that use paired images for training, the datasets (ISTD [6], SRD [12]) consisting of many paired images that are open to the public make it easier to train natural shadow features. In our method, we use these paired images for training.

In this paper, we propose a deep learning-based method that focuses on the correlation of each color channel of shadows. Shadows are strongly influenced by the wavelength of light [1], and the intensity of light that can be acquired
FIGURE 1. Our network is divided into the shadow detection part and the shadow removal part. This structure is based on ST-CGAN [6] and we use the same structure in our previous method [7]. We train this network by using our enhanced dataset, which includes train images with artifacts.

by the camera changes depending on the color of the target. Therefore, in general, it is meaningless to estimate how weak the light intensity of the shadow region is without considering the color of the target. Considering the principle of shadow in nature, each channel should be trained individually so that the physical characteristics of each wavelength can be effectively trained. However, in the image acquisition system of the camera, the image is processed using the information of multiple channels when adjusting the white balance or changing the color space [18]. Since this process mixes the information of each color channel, it is not effective to perform training for each color channel, and it is necessary to consider the correlation of each RGB channel and train them together. Actually, we experimentally found that estimating the RGB color channels separately causes a decrease in the accuracy of shadow removal. In our method, good results were obtained by the newly proposed network structure which considers the correlation between each channel.

In the ISTD and SRD datasets, there is a bias in the situation when collecting training images, and the number of the images is insufficient for general-purpose shadow removal. Therefore, we added some artifacts to the training image based on the shadow model so that the training images could have a more widespread distribution of features. By randomly changing the parameters for adding artifacts during training, we became able to train the network with images with a wide range of features.

Furthermore, by training the network using a loss function focusing on the boundary between the shadow region and the non-shadow region, a natural image can be generated.

The proposed method is based on our previously published method [7]. The overview of our method is shown in Fig. 1. The point that the proposed method uses two networks for shadow detection and shadow removal is the same as our previous method. However, in this method, the dataset for training is processed so that more diverse features can be trained. We also devised the network structure based on theory and experiment to obtain better results. As a result, the proposed method is more versatile than the past method that left good results only for specific images. The training is performed using the ISTD+ dataset [14], and the result from the fine-tuned network using the ISTD [6], SRD [12], and SRD+ [14] datasets are also proposed for comparison with other methods. Our method produces better results in RMSE, PSNR, and SSIM than the state-of-the-art methods for all datasets. The main contributions of our method (Channel Attention GAN: CANet) are summarized below:

- We proposed a new deep learning-based method for shadow removal, focusing on the correlation between each color channel.
- In order to obtain more complex training images, artifacts are added to the training image based on the shadow model.
- By using the loss function that focuses on the boundary of the shadowed area, it is now possible to generate a more natural image.
- For all ISTD, ISTD+, SRD, and SRD+ datasets, our method shows better results than the state-of-the-art methods.

The studies and theories underlying our method are given in section II and III. Section IV describes the details of our proposed method, and section V describes a comparison of experimental results between our method and the other methods.

II. RELATED WORK

A. PHYSICAL MODELING-BASED METHODS

In traditional methods, shadow removal is performed by formulating a physical model of the shadow and performing an inverse operation based on these models [8]–[11]. In these methods, to solve the problem more easily, some parameters of the shadow model are regarded as fixed parameters, and approximated expressions are used. Since the shadow model is not sufficient to represent natural shadows, or the image processing is performed using an approximated shadow model, it has been difficult to accurately restore colors even if the assumed parameters can be accurately estimated. A method that requires user-guided information has been proposed [11], but since giving prior information by the user increases the time and effort of the user and makes it impossible to process a large number of images, shadow removal by a single image is an important task.

B. DEEP LEARNING-BASED METHODS

Shadow removal methods based on deep learning have produced better results than traditional methods. The release of large-scale datasets ISTD [6] and SRD [12], which includes pairs of shadowed and non-shadow images, makes it easier to train networks, and many deep learning-based methods using these datasets have been proposed. Some deep learning-based
method trained with unpaired images was also proposed [15], but there are problems such as some textures being lost in the process of feature translation. Regarding the methods which use paired images for training, Qu et al. [12] proposed a method that predicts the ratio of the amount of change in pixel values in the shadow region (DeshadowNet). Along with this method, they released the SRD dataset consisting of 3088 paired images, which include shadowed and non-shadow images. Wang et al. [6] proposed a method that performs shadow detection and shadow removal in different networks (ST-CGAN). They also released the ISTD dataset consisting of 1870 triplet images, which include shadowed, non-shadow, and shadow mask images. Our proposed method is based on ST-CGAN. Hu et al. [19] proposed a shadow detection method (DSC), focusing on the direction of shadows. In their later study [13], the results of shadow removal using this method are also shown. SID [14] was proposed by Le et al., whose method presumes parameters in a shadow model by using neural networks. In addition, they proposed a dataset called ISTD+ with adjusted color tones of the ground truth images. Cun et al. proposed a method called DHAN [16] that focuses on two layers, the feature layer and the attention layer. They also proposed a network (SMGAN) for generating artificial shadow images in order to train the DHAN with more various images. The shadow removal method proposed by Fu et al. (AEF [17]) generates fusion weight maps for multiple over-exposed images. They also proposed a post-processing network (boundary-aware RefineNet) to reduce artifacts at the boundary between the shadow area and the non-shadow area. We also have already proposed a shadow removal method [7] based on ST-CGAN and the proposed method in this paper is based on our previous method.

III. SUPPORTING METHODS

A. SHADOW MODEL

In this paper, we use the same shadow model as Shor et al [20]. The light intensity $I$ at the position $x$ is described as below:

$$I(x, \lambda) = L(x, \lambda)R(x, \lambda),$$

(1)

where $L$ is the illumination and $R$ is the reflectance. Note that since $I$, $L$, and $R$ depend on the wavelength $\lambda$ and position $x$, they are expressed as a function of $\lambda$ and $x$. Looking at this equation in more detail, it can be divided into two terms related to direct light and indirect light as shown in the following equation.

$$I^{\text{lit}}(x, \lambda) = L^d(x, \lambda)R(x, \lambda) + L^a(x, \lambda)R(x, \lambda),$$

(2)

where $L^d$ is the direct illumination and $L^a$ is the indirect illumination. In the shadowed region, $L^d$ should be 0 because the direct light is blocked by some objects. Therefore, the intensity $I^{\text{shadow}}$ in the shadowed region can be expressed as:

$$I^{\text{shadow}}(x, \lambda) = L^a(x, \lambda)R(x, \lambda).$$

(3)

We use these equations for generating new training images and artifacts. The detail of our method is explained in section IV.

B. IMAGE PROCESSING IN IMAGE ACQUISITION SYSTEM

When the camera acquires an image, various processes are performed on the obtained data. Demosaicing, noise removal, white-balancing, color space conversion, gamma curve application, etc. are performed [18]. For example, the gamma curve application is expressed by the following equation:

$$J_{\text{out}} = J_{\text{in}}^{1/2.2}.$$

(4)

Many of the processes which are performed in the image acquisition system are non-linear processes and can be a problem in image processing based on physical theory. Therefore, when artificial shadow images are generated based on a physical model, more realistic shadow images can be generated by considering the influence of these nonlinear processes.

IV. PROPOSED METHOD

Our method is a learning-based method that removes shadows from a single image. We show that shadows can be effectively removed by the proposed network structure and the extended training dataset.

A. OVERVIEW

Our method is divided into two parts, the shadow detection network $G_D$ and the shadow removal network $G_R$. This structure is based on ST-CGAN [6], and the overview of our method for training is shown in Fig. 2. Since our method is a GAN-based method, discriminator $D_D$ and $D_R$ are defined for the generators $G_D$ and $G_R$. The input to the $G_D$ is only the shadow image, and the output of $G_D$ is the estimated shadow mask. For the discriminator $D_D$, the total of 4 channels of shadow image and shadow mask is the input. On the other hand, the input to the $G_R$ is a total of 6 channels, which includes the shadow image, and the element product of the shadow image and the shadow mask generated by the $G_D$. 

FIGURE 2. The overview of our method for training.
For the discriminator $D_R$, the input data for $G_R$ and the generated image by the $G_R$ are used as the input data. This method is the same as our previous method [7].

B. NETWORK STRUCTURE

The network structure of $G_D$ and $G_R$ is based on Unet++ [24]. It consists of multiple nodes, upsampling, and downsampling. Bilinear upsampling with $2 \times$ magnification is used for the upsampling, and the max pooling layer is used for the downsampling. Each node is a residual block consisting of convolutional layers, batch normalization layers [22], mish activation function [21], and scSE block [23]. Details of the node structure are shown in Fig. 3b.

The shadow detection network $G_D$ has an additional structure after the Unet++ based network to limit the output to the range 0 to 1 (Fig. 3a). In $G_R$, which is the shadow removal network, ColorBlock (Fig. 3c) that weights each color channel is added after the Unet++ based network. $G_R$ is designed to learn the difference between the input image and the ground truth image.

Since $G_R$ is a network that estimates the difference between the input image and the ground truth image, the content $\Delta$ that $G_R$ should estimate can be written as follows based on Eq.2, 3:

$$\Delta = l_{gt} - l_{input} = P(l_{lit}(x, \lambda)) - P(l_{shadow}(x, \lambda)) \approx l_{lit}(x, \lambda) - I_{shadow}(x, \lambda).$$

The function $P$ indicates the image processing during the camera’s image acquisition system. For the sake of simplicity, we do not consider $P$ in this sub-section. Since the image has only three color channels composed of R, G, and B, it can be approximated that $L_d(x, \lambda)$ of Eq.5 is not a function of $\lambda$ but a function of R, G, and B. In addition, assuming that the intensity of the direct light from the light source is constant for the entire image, $L_d(x, \lambda)$ will not be a function of $x$ but will be a constant value in each color channel. From the above, Eq.5 can be approximated as Eq.6:

$$\Delta \approx L_d(x, \lambda)R(x, \lambda) \approx \left( \frac{L^d_R(x)}{R_R(x)} \right) \left( \frac{L^d_G(x)}{R_G(x)} \right) \left( \frac{L^d_B(x)}{R_R(x)} \right).$$

Since the ColorBlock has the role of weighting each color channel with a specific value as the final process, it can be interpreted as estimating $L^d_R$, $L^d_G$, and $L^d_B$ for each image. By estimating the weight of each channel using the fully connected layer, the value of $L^d_{R,G,B}$ can be estimated while being influenced by each other channel. ColorBlock is effective for training the relationships between color channels because the convolutional layer cannot efficiently train the relationships between channels. By training $L^d_{R,G,B}$ and $R_{R,G,B}(x)$ individually, it is expected that our network is trained as a robust estimation network for changes in image tint.

The discriminator consists of convolutional layers, batch normalization layers, and Mish layers. These layers are connected in sequence. By setting the stride of the convolutional layer to 2 for every two convolutional layers, the size of the final output is $16 \times 16$. The L2 loss is used to calculate the loss of the discriminator.

C. LOSS FUNCTIONS

1) GENERATOR $G_D$

Generator $G_D$ is the network for shadow detection. The shadowed image $l_{input}$ consists of 3 RGB channels input.
The shadow mask $M_{est}$ can be estimated as follows:

$$M_{est} = G_D(I_{input}; \theta_{G_D}),$$

(7)

where $\theta_{G_D}$ is the weight of the generator $G_D$. In order to train $G_D$, we have to minimize the loss function $L_{G_D}(\theta_{G_D})$, which is composed of $L_{MSE}^{G_D}$, $L_{adv}^{G_D}$, and $L_{feat}^{G_D}$.

$L_{MSE}$ is the pixel loss and calculates the L2 norm between two images. Pixel loss calculates the difference between the generated shadow mask $M_{est}$ and the ground truth shadow mask $M_{gt}$ on a pixel-by-pixel basis. Applying MSE loss when using GAN is useful to prevent the gradient vanishing problem [25]. $L_{G_D}$ is expressed as:

$$L_{MSE}^{G_D} = \sum_i ||M_{gt} - M_{est}||_2,$$

(8)

where $N$ is the batch size.

$L_{G_D}^{adv}$ is the adversarial loss calculated using the discriminator $D_D$. Networks trained using only pixel-by-pixel MSE loss are trained towards an average solution across the dataset [26], which is likely to produce blurred or unnatural images. To solve this problem, a network structure called Generative Adversarial Networks (GAN) consisting of a generator and a discriminator was proposed [27]. Adversarial loss is used to train these networks. By applying this loss, it becomes possible to train the network to generate a natural image that more closely matches the image distribution of the training data set. $L_{G_D}^{adv}$ is expressed as:

$$L_{G_D}^{adv} = \sum_i ||V_{real} - D_D(I_{input}, M_{est}; \theta_{D_D})||_2,$$

(9)

where $V_{real}$ is a random valued matrix that follows Gaussian distribution with an average of 0.5.

$L_{G_D}^{feat}$ is a feature loss calculated using our shadow removal network $G_R$. The purpose of applying this loss is to adjust the output of $G_D$ so that $G_R$ can generate more realistic images. The best shadow removal performance may not always be achieved when the ground truth mask image is used as the input to $G_R$. This is because the ground truth mask image does not always represent the shadow part correctly and might be misaligned. Therefore, instead of training the two networks completely separately, we train them while influencing each other. $L_{G_D}^{feat}$ is expressed as:

$$L_{G_D}^{feat} = \sum_i ||I_{gt} - G_R(I_{input}, I_{input} \odot M_{est}; \theta_{G_R})||_2,$$

(10)

where $\theta_{G_R}$ is the weight of $G_R$ and it is fixed while training $G_D$.

Using the Eq.8-10, the loss function $L_{G_D}(\theta_{G_D})$ is expressed as follows:

$$L_{G_D}(\theta_{G_D}) = L_{G_D}^{MSE} + \frac{10}{b + 1} \times L_{G_D}^{adv} + 10 \times L_{G_D}^{feat},$$

(11)

where $b$ is the number of the batches trained. If the influence of $L_{G_D}^{adv}$ is too large, it may lead to a decrease in the RMSE of the final output image, so $L_{G_D}^{adv}$ is set to reduce the weight as the training progresses. These weights are determined experimentally.

2) GENERATOR $G_R$

Generator $G_R$ is the network for shadow removal. The input data is 6-channel data obtained by concatenating the input image and the additional image given by multiplying the input image and the estimated shadow mask. The shadow-removed image is generated as follows:

$$I_{output} = G_R(I_{input}, I_{input} \odot M_{est}; \theta_{G_R}).$$

(12)

Similar to $G_D$, the loss function $L_{G_R}(\theta_{G_R})$ needs to be minimized to train the generator $G_R$. The loss function $L_{G_R}(\theta_{G_R})$ consists of three loss functions $L_{MSE}^{G_R}$, $L_{adv}^{G_R}$, and $L_{edge}^{G_R}$.

$L_{MSE}^{G_R}$, $L_{adv}^{G_R}$, and $L_{edge}^{G_R}$ are applied in $L_{G_R}(\theta_{G_R})$ for the same reason as described in section IV-C1. They are expressed by the following formula:

$$L_{MSE}^{G_R} = \sum_i ||I_{gt} - I_{output}||_2$$

(13)

$$L_{adv}^{G_R} = \sum_i ||V_{real} - D_R(I_{input}, I_{input} \odot M_{est}, I_{output}; \theta_{D_R})||_2.$$

(14)
where $D_R$ is the discriminator for $G_R$. The difference between $L_{G_R}(\theta_{G_R})$ and $L_{D_R}(\theta_{D_R})$ is the adoption of a new loss function $L_{edge}^{G_R}$:

$$L_{edge}^{G_R} = \sum_i ||(I_{gt} - I_{output}) \circ M_{edge}||_2.$$  

Applying $L_{edge}^{G_R}$ made the boundary between the shadow area and the non-shadow area natural, which greatly contributed to the improvement of shadow removal performance. The effectiveness of $L_{edge}^{G_R}$ is described in section V-C.

Using the Eq.12-16, the loss function $L_{G_R}(\theta_{G_R})$ is expressed as follows:

$$L_{G_R}(\theta_{G_R}) = L_{MSE}^{G_R} + \frac{1}{10 \times \sqrt{b} + 1} \times L_{adv}^{G_R} + 25 \times L_{edge}^{G_R},$$  

where $b$ is the number of the batches trained. We set these weights experimentally.

3) DISCRIMINATORS

Discriminators $D_D$ and $D_R$ are one of the components of GAN for calculating adversarial loss. As mentioned in section IV-B, L2 loss is used for training the discriminators. The loss function $L_{D_D}(\theta_{D_D})$ for discriminator $D_D$ and $L_{D_R}(\theta_{D_R})$ for $D_R$ are calculated as follows:

$$L_{D_D}(\theta_{D_D}) = \sum_i ||V_{fake} - D_D(I_{input}, M_{est}; \theta_{D_D})||_2$$

and

$$L_{D_R}(\theta_{D_R}) = \sum_i ||V_{real} - D_R(I_{input}, I_{input} \circ M_{gt}, I_{gt}; \theta_{D_R})||_2 + \sum_i ||V_{fake} - D_R(I_{input}, I_{input} \circ M_{est}, I_{output}; \theta_{D_R})||_2/2.$$

$V_{real}$ is a random valued matrix that follows Gaussian distribution with an average of 0.5 and $V_{fake}$ is the same as $V_{real}$ but its average is $-0.5$. We train our discriminator $D_D$ and $D_R$ by minimizing $L_{D_D}(\theta_{D_D})$ and $L_{D_R}(\theta_{D_R})$.

D. TRAINING DATASET

We used the ISTD+ [14] dataset to train our proposed network. The ISTD+ dataset is a color-corrected dataset, based on the ISTD dataset [6]. The ISTD+ dataset includes 1870 triplet images, and it is divided into 1330 triplet training images and 540 triplet test images. In our method, all images were randomly processed at each epoch in order to increase the complexity of the training dataset. We applied flipping, rotating, scaling and clipping, changing colors, and inserting artifacts. When applying changing colors and inserting artifacts, we propose a new method based on the camera’s image acquisition system [18] and shadow model so that we can generate more natural images. The execution result of each image processing is shown in Fig. 5. The effectiveness of this enhanced training dataset is shown in Sec. V-C.

1) FLIPPING

Flipping is the process of flipping an image horizontally or vertically. This process can be represented by the following equation:

$$J_{out}(i, j) = \begin{cases} J_{in}(i, n - j) & \text{Horizontal flip} \\ J_{in}(m - i, j) & \text{Vertical flip} \end{cases},$$  

where $m$ is the height and $n$ is the width of the image $J_{in}$. The input image during network training is flipped horizontally with a 50% probability and then flipped vertically with a 50% probability.

2) ROTATING

In our method, the training image is rotated by 0, 90, 180, and 270 degrees with the same probability. The rotation can be expressed by the following equation:

$$J_{out}(i, j) = \begin{cases} J_{in}(i, j) & \text{Rotating 0 degrees.} \\ J_{in}(j, m - i) & \text{Rotating 90 degrees.} \\ J_{in}(m - i, n - j) & \text{Rotating 180 degrees.} \\ J_{in}(n - j, i) & \text{Rotating 270 degrees.} \end{cases}.$$
3) SCALING AND CLIPPING
The image is enlarged by a factor of 2 using the Nearest neighbor method and clipped at random positions to be the same size as the original image. This process is applied with a probability of 33% during the training.

4) CHANGING COLORS AND INSERTING ARTIFACTS
In order to increase the variety of the training dataset, it is necessary to acquire images that cover various situations such as many kinds of target objects and light sources. However, it is quite time-consuming and difficult to collect images by changing the target object and the light source in the real world. Therefore, in our method, we modified the features of the captured object based on the shadow model to increase the diversity of the training dataset.

From Eq.1, we can understand that $R(x, \lambda)$ changes when the target objects change. Since $L(x, \lambda)$ is a parameter related to direct light and indirect light, it is not affected by the target object. From Eq.2 and 3, the ratio of the light intensity in the shadow region between the shadowed image $I_{input}$ and the non-shadow image $I_{gt}$ can be expressed as follows:

$$\frac{I_{shadow}}{I_{input}} = \frac{L^a(x, \lambda)}{L^d(x, \lambda) + L^a(x, \lambda)} . \quad (22)$$

From the above equation, we can see that the ratio of the light intensity in the shadow region does not depend on the type of the target object, but depends on the parameters related to the light falling on the object. Therefore, if we consider that the object has been changed while the light parameters remain the same, we can generate a new shadow region using the following equation:

$$I_{shadow_{input\_new}} = \frac{L^a(x, \lambda)}{L^d(x, \lambda) + L^a(x, \lambda)} \times I_{shadow_{gt\_new}} . \quad (23)$$

In the ideal state, the non-shadow area should show the same value for $I_{input}$ and $I_{gt}$, so the whole image $I_{input\_new}$ can be expressed as follows by applying Eq.23.

$$I_{input\_new} = I_{shadow_{input\_new}} \cup I_{non-shadow_{input\_new}} = I_{shadow_{gt\_new}} \cup I_{non-shadow_{gt\_new}}$$

$$I_{shadow_{gt\_new}} = I_{shadow_{gt}} \cup I_{non-shadow_{gt\_new}} \cup I_{non-shadow_{gt\_new}} \times I_{input_{gt\_new}}$$

$$I_{input_{gt\_new}} = I_{input_{gt}} \cup I_{gt\_new}, \quad (24)$$

where $I_{gt\_new}$ is the new ground truth light intensity. Our new $I_{gt\_new}$ is based on $I_{gt}$. When creating $I_{gt\_new}$, each color channel was multiplied by a random coefficient, and the color of a part of the image was randomly changed to add an artifact. The artifacts are incorporated to train the situation in which there are many objects in the shadow region.

It should be noted that Eq.24 is an equation expressing the intensity of light. The images we handle are converted in a nonlinear color space to make them look natural to the human eyes [18]. Therefore, when image processing based on the shadow model is performed in a non-linear color space, an unnatural shadow image may be generated. In our method, to reduce the unnaturalness of the synthetic image generated in an inappropriate color space, the color space is converted before the image processing is performed. Specifically, before generating an image based on Eq.24, the color space of the image is converted using the following equation:

$$J_{out} = \begin{cases} \frac{J_{in}}{12.92} & (J_{in} \leq 0.04045) \\ ((J_{in} + 0.055)/1.055)^{2.4} & (J_{in} > 0.04045) \end{cases} . \quad (25)$$

After performing image processing in a new color space, reverse processing is performed to return it to the original color space. This series of processing is applied with a probability of 60% during the training.

E. ROTATE AVERAGING PROCESS
Since our network is not rotation-invariant, the results may change as the rotated images are processed. Our method
uses a rotation averaging process that averages multiple output images generated from multiple rotated input images to achieve improved results. We have used the same process in our previous method [28], and this process can be used as a general-purpose process.

**F. IMPLEMENTATION AND TRAINING**

Our method is implemented using Python and PyTorch [29]. The training is done using the ISTD+ dataset [14]. The proposed networks are optimized using Adam optimizer [30]. The input image has been resized to 256 x 256 with nearest-neighbor interpolation and has been randomly pre-applied with image processing as shown in Sec.IV-D. The batch size is set to 4 and the initial learning rate of the Adam optimizer is set to 0.0002. Each network is trained by 3000 epochs, and the weight with the lowest RMSE for the test data is used as the result of this paper.

In addition, fine-tuning [31] was performed using multiple data sets for comparison with other methods. The datasets used for fine-tuning are ISTD [6] (including 1330 training images), SRD [12] (including 2680 training images), and SRD+ (including 2680 training images). The image adjusting method [14] proposed by Le et al. is also applied to the SRD dataset to create the SRD+ dataset. For fine-tuning, the dataset was changed and trained for 100 additional epochs, and the weights with the best results were used as the results in this paper.

**V. EXPERIMENTAL RESULTS**

**A. EVALUATION METRIC**

We use RMSE in the Lab color space for quantitative comparisons, similar to the conventional methods. As SID author Le [14] mentions on their project page, the evaluation code used in the traditional methods including SID [14] calculates MAE instead of RMSE. Therefore, in this paper, instead of comparing with the numerical values described in the papers of the conventional methods, the results were recalculated using the experimental result images published by the authors of each method. For the color space conversion, we used the rgb2lab function available in Python’s scikit-image, which converts sRGB colorspace (IEC 61966-2-1: 1999) to CIE Lab colorspace. In the Lab color space, the range of values that can be taken for each channel is different. Since each Lab channel shows a different dimension of information, our paper shows the value obtained by calculating RMSE for each channel and finally adding them together. This paper also describes the results of comparisons using PSNR [32] and SSIM [32] in the non-converted sRGB color space. The code for quantitative comparison is available on our GitHub page (https://github.com/ryo-abiko/CANet).

**B. SHADOW REMOVAL EVALUATIONS**

In this section, we compare our proposed method (CANet) with other notable methods, including Yang [9], Guo [10], Gong [11], ST-CGAN [6], Nagae [7], DSC [19], SID [14], DHAN [16], AEF [17].

Quantitative evaluation is performed using ISTD, ISTD+, SRD, and SRD+ datasets. As in the conventional methods [16], [17], individual network weights are prepared for each dataset to evaluate the shadow removal performance. In this paper, the comparison method is quantitatively evaluated using the shadow removal result images published by the authors of each paper. Therefore, the shadow removal results which are not open to the public are not compared in this paper. For the results of Yang [9], Guo [10], and Gong [11], we used the results published by the authors of ST-CGAN [6].

The results are shown in Table 3 for the ISTD+ dataset, Table 4 for the ISTD dataset, Table 5 for the SRD dataset, and Table 6 for the SRD+ dataset.

Since our method is trained using the ISTD+ dataset, we first check the results in Table 3. Table 3 shows that our method performed best for RMSE in the Lab color space and also for SSIM in the sRGB color space. Also from Table 4, 5, and 6, we can see that our method outperforms the conventional method in all categories for the ISTD, SRD, and SRD+ datasets. To allow for easy comparison, Table 2 shows the average value of the results from the ISTD, ISTD+, and SRD datasets. SRD+ is not used for this statistic because the author of AEF [17] did not publish the results from them. This result shows that our method does not have good performance only for a specific dataset, but can be generically applied for various datasets.

For qualitative evaluation, we use the ISTD+ and SRD+ datasets. The results for ISTD+ are shown in Fig. 6 and the results for SRD+ are shown in Fig. 7. From Fig. 6, we can see that our method can remove shadows even when there are areas where the color is monotonous or where the color changes significantly in the shadow regions. This result shows the effect of training on a dataset that has been extended to include a variety of colors and artifacts. In addition, even in a smooth region, the boundary between the shadow region and the non-shadow region is not noticeable. It is considered that this is because the training was performed using the edge loss, and the network was focused on the boundary of the shadow region. Since our method generates a shadow mask once, it is possible to clarify the area where
the shadow should be removed. From Fig. 7, we can see the advantage of the structure of our network, and we can confirm that shadow removal can be done with little color change in the non-shadow areas.

C. EFFECTIVENESS OF OUR NEWLY PROPOSED COMPONENTS

In this paper, we mainly proposed three new elements. (1) We adopted a structure called ColorBlock to effectively
train the correlation between color channels. (2) To increase the complexity of the training dataset, the color of the training image was randomly changed and artifacts are added. (3) We introduced Edge Loss, which is a loss function focusing on the edge of the shadow region.

In order to confirm the effectiveness of these new proposals, the results without applying these proposals are shown in Table 7. The transition in shadow removal performance for the test dataset during the training is also shown in Fig. 8. The horizontal axis in Fig. 8 shows the number of epochs of training, and the vertical axis shows the minimum RMSE of the shadow removal results for the test dataset at that point. As can be seen from Table 7, the results are improved when each method is applied. In particular, we can see that the enhancement of the training dataset has a significant effect. Fig. 8 also shows that when the training dataset was not extended, the network was overtrained and did not produce good results.

**D. RUNNING TIME**

We used different environments for network training and evaluation. It is because the GeForce RTX 2080 Ti GPU ran out of memory to train our proposed networks in batch.
However, the qualitative evaluation shows that our method performs better than the comparison methods, indicating that the shadow removal is sufficient.

VI. CONCLUSION

In this paper, we proposed a new method called CANet for single-image shadow removal. Our method is a combination of a network that detects shadows and removes shadows. By introducing ColorBlock, which can learn the correlation between each channel effectively, the accuracy of final shadow removal images has been improved. In addition, in order to increase the complexity of the training data set, the color of the training image was changed and artifacts were added while considering the shadow model and the image acquisition system of the camera. By using Edge Loss, which focuses on the edges of the shadow region, we have achieved natural shadow removal even in regions with little texture. Since our method uses two Unet++ based networks, the number of parameters is large. However, within the scope of our study, this is the network structure that gives the best performance. As a result, our method outperforms existing methods in multiple datasets, both quantitatively and qualitatively.

REFERENCES

[1] J. Tian, J. Sun, and Y. Tang, “Tricolor attenuation model for shadow detection,” IEEE Trans. Image Process., vol. 18, no. 10, pp. 2355–2363, Oct. 2009.
[2] S.-I. Choi, C. Kim, and C.-H. Choi, “Shadow compensation in 2D images for face recognition,” Pattern Recognit., vol. 40, no. 7, pp. 2118–2125, Jul. 2007.
[3] W. Zhang, X. Zhao, J.-M. Morvan, and L. Chen, “Improving shadow suppression for illumination robust face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 3, pp. 611–624, Mar. 2019.
[4] N. Friedman and S. Russell, “Image segmentation in video sequences: A probabilistic approach,” 2013, arXiv:1302.1539.
[5] A. Sanin, C. Sanderson, and B. C. Lovell, “Improved shadow removal for robust person tracking in surveillance scenarios,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 141–144.
[6] J. Wang, X. Li, and J. Yang, “Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1788–1797.
[7] T. Nagae, R. Abiko, T. Yamaguchi, and M. Ikehara, “Shadow detection and removal using GAN,” in Proc. 28th Eur. Signal Process. Conf. (EUSIPCO), Jan. 2021, pp. 630–634.
[8] E. Arbel and H. Hel-Or, “Shadow removal using intensity surfaces and texture anchor points,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 6, pp. 1202–1216, Jun. 2011.
[9] Q. Yang, K.-H. Tan, and N. Abuja, “Shadow removal using bilateral filtering,” IEEE Trans. Image Process., vol. 21, no. 10, pp. 4361–4368, Oct. 2012.
[10] R. Guo, Q. Dai, and D. Hoiem, “Paired regions for shadow detection and removal,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2956–2967, Dec. 2013.
[11] H. Gong and D. Cosker, “Interactive shadow removal and ground truth for variable scene categories,” in Proc. Brit. Mach. Vis. Conf., 2014, pp. 1–11.
[12] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau, “DeshadowNet: A multi-context embedding deep network for shadow removal,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4067–4075.
[13] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng, “Direction-aware spatial context features for shadow detection and removal,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2795–2808, Nov. 2020.
[14] H. Le and D. Samaras, “Shadow removal via shadow image decomposition,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8578–8587.
[15] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, “Mask-ShadowGAN: Learning to remove shadows from unpaired data,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2472–2481.

[16] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, “Mask-ShadowGAN: Learning to remove shadows from unpaired data,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2472–2481.

[17] X. Cun, C.-M. Pun, and C. Shi, “Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN,” in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 10680–10687.

[18] L. Fu, C. Zhou, Q. Guo, F. Jaebei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang, “Auto-exposure fusion for single-image shadow removal,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10571–10580.

[19] H. C. Karaimer and M. S. Brown, “A software platform for manipulating the camera imaging pipeline,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 429–444.

[20] X. Hu, L. Zhu, C.-W. Fu, J. Qin, and P.-A. Heng, “Direction-aware spatial context features for shadow detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7454–7462.

[21] Y. Shor and D. Lischinski, “The shadow meets the mask: Pyramid-based shadow removal,” Comput. Graph. Forum, vol. 27, no. 2, pp. 577–586, 2008.

[22] D. Misra, “Mish: A self regularized non-monotonic activation function,” 2019, arXiv:1908.08681.

[23] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.

[24] A. G. Roy, N. Navab, and C. Wachinger, “Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2018, pp. 421–429.

[25] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Cham, Switzerland: Springer, 2018, pp. 3–11.

[26] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2794–2802.

[27] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4681–4690.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[29] R. Abiko and M. Ikehara, “Single image reflection removal based on GAN with gradient constraint,” IEEE Access, vol. 7, pp. 148790–148799, 2019.

[30] A. Paszke, S. Gross, S. Chintal, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in Proc. NIPS Workshop Autodiff, 2017, pp. 1–4. [Online]. Available: https://openreview.net/forum?id=BJJsmrCZ

[31] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” 2014, arXiv:1412.6980.

[32] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1299–1312, May 2016.

[33] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

**RYO ABIKO** received the B.E. and M.E. degrees in electrical engineering from Keio University, Yokohama, Japan, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree under the supervision of Prof. M. Ikehara. His research interests include image filtering, denoising, and image processing using deep learning.

**MASAAKI IKEHARA** (Senior Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1984, 1986, and 1989, respectively. He was a Lecturer at Nagasaki University, Nagasaki, Japan, from 1989 to 1992. In 1992, he joined the Faculty of Engineering, Keio University. From 1996 to 1998, he was a Visiting Researcher at the University of Wisconsin–Madison, Madison, WI, USA, and Boston University, Boston, MA, USA. He is currently a Full Professor with the Department of Electronics and Electrical Engineering, Keio University. His research interests include multi-rate signal processing, wavelet image coding, and filter design problems.