IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 6, 2019, accepted November 21, 2019, date of publication November 26, 2019,
date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955994

Single Image Reflection Removal Using Non-Linearly Synthesized Glass Images and Semantic Context

BYEONG-JU HAN AND JAE-YOUNG SIM, (Member, IEEE)
School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea

Corresponding author: Jae-Young Sim (jysim@unist.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) within the Ministry of Science and ICT (MSIT) under Grant 2017R1A2B4011970.

ABSTRACT An image captured through a glass plane usually contains both of a target transmitted scene behind the glass plane and a reflected scene in front of the glass plane. We propose a semantic context based network to remove reflection artifacts from a single glass image. We first investigate a non-linear intensity mapping relationship for glass images to synthesize more realistic training sets. Then we devise an efficient reflection removal network using multi-scale generators and an interpreter, where the semantic context of the transmission image is adopted as a high level cue for the interpreter to guide the generators. We also provide a new test data set of real glass images including the ground truth transmission and reflection images. Experiments are performed on four test data sets and we show that the proposed algorithm decomposes an input glass image into a transmission image and a reflection image more faithfully compared with the four existing state-of-the-art methods.

INDEX TERMS Reflection removal, image restoration, deep learning, semantic context.

I. INTRODUCTION

Glass material has been used in many places, for example, glass display windows in retail shops are used to show products to customers while protecting the products. However, when taking pictures through glass, light reflection occurs on glass planes, which reduces the visibility of target transmitted scenes behind the glass planes and thus degrades the performance of computer vision techniques such as text recognition [1], object detection [2], and semantic segmentation [3]. From the literatures, lots of attempts have been made to reconstruct transmission images faithfully by removing reflection artifacts from glass images.

Many existing methods use multiple glass images taken under different capturing conditions to extract the characteristics of typical reflection images. However, special capturing environments are often required such as replacing camera filters or adjusting focal lengths, which cause limitations to practical application. Single image based reflection removal techniques employ unique characteristics of reflection images such as smoothness prior [4] or ghosting cue [5], however, they often fail to work on glass images with diverse characteristics. Recently, deep learning based techniques have been proposed and outperformed the existing methods, but they also have problems to obtain sufficient training data.

Most of the existing methods regard a glass image as the sum of a transmission image and a reflection image. Such a linear model simplifies the objective functions to separate transmission image and reflection image by applying the conventional solvers such as Alternating Direction Minimizing (ADM) [6]. Deep learning based methods have been improving the performance of reflection removal, however, they also employ the simple linear model of glass image to generate synthetic training sets. In this paper, we investigate a non-linear intensity mapping relationship to synthesize more realistic glass images, which estimates the most probable pixel value of glass image when a pair of pixel values for transmission image and reflection image are given. Moreover, most of the existing supervised learning methods train deep networks by simply comparing the pixel values between a generated image and its ground truth, which often degrades the continuity of spatial context over an entire image area. We also devise a novel reflection removal algorithm using semantic

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

context that encourages the restored transmission image to have the same semantic information to that of the ground truth transmission image. Specifically, we adopt a semantic segmentation network that provides prediction maps where the pixel values in each map represent the relevance to a certain class. Then we compare not only the pixel values but also the prediction maps between the generated transmission image and the ground truth image. Experimental results show that the proposed single image reflection removal algorithm restores both of the transmission image and the reflection image faithfully and outperforms the existing state-of-the-art methods.

The contributions of this paper are summarized as follows.

1) We empirically derived a non-linear intensity mapping relationship from real glass images to synthesize more realistic training images.
2) We adopted a high level feature of semantic context in an entire image to train a reflection removal network more reliably.
3) We provided a new data set of real glass images including 71 triplets of glass image, transmission image and reflection image. We also suggested a robust methodology to evaluate the performance of reflection removal.

The remaining of this paper is organized as follows. Section II introduces the related work. Section III explains the proposed algorithm. Section IV shows the experimental results. Section V concludes the paper with future research.

II. RELATED WORK

The existing methods using multiple glass images were based on distinct behaviors of the reflection images observed across the multiple input glass images. At a fixed camera position, multiple glass images were taken by varying camera settings, such as focal length [7], flash [8], [9], and polarization [10]. Also, multiple glass images taken at different camera positions were used to estimate different depth maps for reflected scene and transmitted scene [11]. Xue *et al.* [12] computed optical flow maps where each pixel represents the movement of the transmitted scene or the reflected scene. Han and Sim [13] accommodated the consistency of transmission gradients across the multiple glass images using a low rank matrix completion scheme.

The single image based reflection removal methods characterized the intrinsic properties observed in reflection images. Li and Brown [4] separated a reflection image from a glass image according to the smoothness prior that reflection images usually yield narrow distribution of gradient magnitude compared to transmission images. Wan *et al.* [14] computed the degree of blurriness at each pixel in multiple scales and suppressed the gradients at pixels with high blurriness. Shih *et al.* [5] removed the ghosting artifacts caused by double reflection on both of the front and back sides of glass plane. Levin *et al.* [15] minimized the number of X-junction in the glass image to separate reflection and transmission images from each other.

In recent years, deep learning based methods have been achieving improved performance of reflection removal. Fan *et al.* [16] proposed a two-step network where the first generator estimates the gradient map of the transmission image from an input glass image, and the second generator serially connected to the first generator reconstructs a transmission image from the estimated gradient map. Wan *et al.* [17] suggested to connect the two generators in parallel, where the generator estimating a gradient map guides the other generator to provide a correct transmission image through connections between intermediate feature maps of the two generators. Zhang *et al.* [18] defined a feature loss which compares the feature maps of a generated image and its ground truth image, and trained the characteristics of transmission images using adversarial learning. Yang *et al.* [19] designed a bi-directional imaging network which roughly predicts a reflection image from an initially estimated transmission image, and then predicts a detail transmission image again from the predicted reflection image.

Unlike the existing deep learning based methods, the proposed algorithm compares low-level features of pixel values as well as high-level features of semantic context between the generated transmission image and the ground truth transmission image. Moreover, whereas Zhang *et al.* [18] also defined a loss term using feature maps and designed an image recognition network predicting the class where the input image belongs, the proposed algorithm utilizes the feature maps obtained by a semantic segmentation network that provides the spatial context of target transmitted scene by computing the relevance of image regions to each class.

III. PROPOSED METHOD

We propose a single image reflection removal network in this paper. We first synthesize a training set of realistic glass images by investigating a non-linear relationship of pixel values in real glass images. Then we design multi-scale generators to estimate the transmission image as well as the reflection image from an input glass image, which are guided to provide reliable results by an interpreter based on the semantic context of target transmission images.

A. GLASS IMAGE SYNTHESIS

In general, a glass image G is linearly modeled as the sum of a transmission image T and a reflection image R such that

$$G(\mathbf{p}) = T(\mathbf{p}) + R(\mathbf{p}) \tag{1}$$

where $\mathbf{p}$ is a pixel location. The existing methods of single image reflection removal usually adopt this linear model to generate synthetic data sets for training. For example, Fan *et al.* [16] randomly selected two images for a transmission image and a reflection image, respectively, and synthesized a glass image based on the linear model in Eq. (1) after modifying the reflection image to become blurred and dark. However, as shown in Fig. 1, the linear model does not completely describe the relationship among G, T and R. Fig. 1 (a) shows real glass images and Fig. 1 (b) exhibits the

(a) (b) (c)

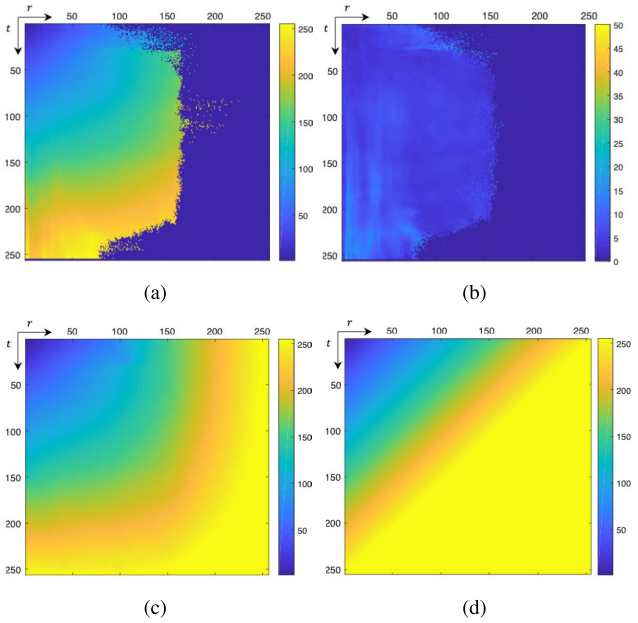
FIGURE 1. Limitation of linear glass image model. (a) Real glass images. (b) Ground truth transmission images captured without glass plane. (c) The estimated reflection images which are obtained by subtracting the ground truth transmission images from the glass images.

associated transmission images which are taken by removing a glass plane in front of a camera. Fig. 1 (c) shows the estimated reflection images which are obtained by subtracting the transmission images from real glass images. We see that the estimated reflection images still exhibit strong scene structures of the transmission images, which shows the limitation of the conventional linear modeling for glass images.

To synthesize more realistic glass images, we empirically derive a non-linear intensity mapping relationship which associates an intensity of the glass image with a given pair of the intensities of the transmission image and the reflection image. In practice, we collect 537 real image sets from [17] where each image set is composed of a glass image, a transmission image, and a reflection image. We observe the combination of three pixel values $(G_i(\mathbf{p}), T_i(\mathbf{p}), R_i(\mathbf{p}))$ at a pixel location $\mathbf{p}$ in each i-th image set, and we gather all the combinations at every pixel location over all image sets to define $\Phi(t, r)$, the set of $G_i(\mathbf{p})$'s when $T_i(\mathbf{p}) = t$ and $R_i(\mathbf{p}) = r$.

$$\Phi(t, r) = \{G_i(\mathbf{p}) | T_i(\mathbf{p}) = t \text{ and } R_i(\mathbf{p}) = r, \forall i \text{ and } \mathbf{p}\}. \quad (2)$$

Figs. 2 (a) and (b) visualize the average and the standard deviation of the elements in $\Phi(t, r)$ associated with t on the y-axis and r on the x-axis, respectively. We exclude unreliable points of (t, r) where the corresponding $\Phi(t, r)$ has less than 100 elements or yields a standard deviation of elements larger than 8. We observe that the average of elements in $\Phi(t, r)$ is not linearly proportional to t or r, and the variation of elements in $\Phi(t, r)$ is relatively small. Then we generate a complete intensity mapping $\mathcal{M}$, as shown in Fig. 2 (c), by filling the undetermined regions in Fig. 2 (a) using an in-painting method [20]. We see that, according to $\mathcal{M}$, the intensities of the glass image are not linearly related with that of the transmission image and the reflection image. For comparison, we also visualize the linear intensity mapping model in Fig. 2(d) where the intensities larger than 255 are cropped to 255. We see that the linear glass image model does not describe the characteristics of real glass images completely.



(a) (b) (c) (d)

FIGURE 2. Relationship of intensity values in real glass images. (a) The average of elements in $\Phi(t, r)$. (b) The standard deviation of elements in $\Phi(t, r)$. (c) The resulting complete non-linear intensity mapping model. (d) The linear intensity mapping model.

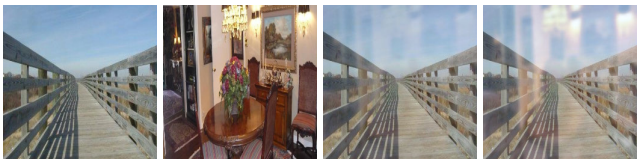
Algorithm 1 Glass Image Generation

```
procedure Synthesize(T, R)
    α ← U(0.2, 1.0), β ← U(0, 5)
    R̄ ← Gaussianblur(R, β)
    R̃ ← αR̄
    for all p in an image do
        G(p) ← M(T(p), R̃(p))
    end for
    return G
end procedure
```

We use the derived non-linear intensity mapping $\mathcal{M}$ to generate synthetic glass images. We randomly select two images from ADE20K dataset [21] for a transmission image and a reflection image, respectively, followed by blurring and attenuating the reflection image with randomly selected parameters to reflect various characteristics of reflection images. Then, at each pixel $\mathbf{p}$, we read the pair of intensity values of the transmission image T and the processed reflection image $\tilde{R}$, which are used to synthesize the intensity value for the glass image as

$$G(\mathbf{p}) = \mathcal{M}(T(\mathbf{p}), \tilde{R}(\mathbf{p})). \quad (3)$$

The proposed method for synthetic glass image generation is summarized in Algorithm 1, where $\mathcal{U}(a, b)$ denotes an uniform random distributor from a to b. For training, we generate 30K synthetic glass images in total with randomly chosen α and β which control the attenuation and blurriness of reflection images, respectively. Fig. 3 shows the examples of



(a) (b) (c) (d)

FIGURE 3. Synthesized glass images. (a) A transmission image. (b) A reflection image. The glass images synthesized by the proposed non-linear intensity mapping with (c) $\alpha = 0.6$ and (d) $\alpha = 0.9$, respectively.

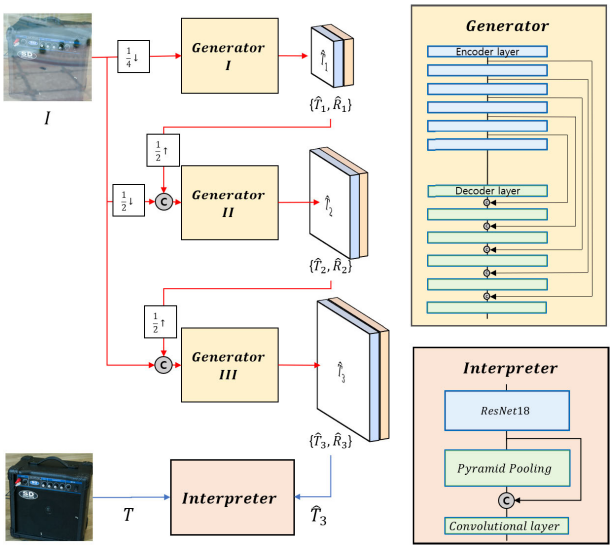


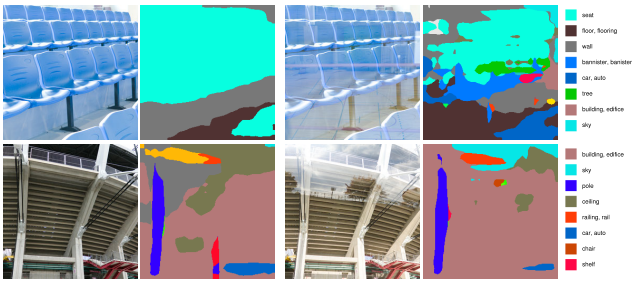
FIGURE 4. Proposed network architecture for reflection removal.

synthesized glass images with different values of the attenuation parameter α.

B. REFLECTION REMOVAL NETWORK

The proposed network architecture consists of multi-scale *generator*s and an *interpreter* as shown in Fig. 4. Each generator at a certain image scale k predicts not only the transmission image $\hat{T}_k$ but also the reflection image $\hat{R}_k$, which are then upsampled and concatenated to the input glass image to become the input to the serially connected generator at the finer image scale $k+1$. We employ three generators that have the same network architecture and provide the resulting images at the scales of 0.25, 0.5 and 1.0 to the original input glass image, respectively. Inspired by UNet [22], we also use the network architecture with encoder and decoder which are connected to each other by skip connections. Each layer includes two convolutional layers followed by batch normalization and ReLU activation. Note that we resize the intermediate feature maps by setting the strides of the two convolutional layers to 2 and 1, respectively.

However, a generative network that is trained to minimize the color difference between a predicted image and the ground truth image often yields the color distortion in local image regions, since the predicted optimal colors may not completely reflect the semantic context of the ground truth image. To overcome this limitation, we additionally employ the interpreter which guides the generators to be



(a) (b)

FIGURE 5. Semantic segmentation results. (a) Transmission images. (b) Glass images.

trained reliably by comparing the semantic contexts between the resulting transmission image and the ground truth image. Specifically, we use a semantic segmentation network as the interpreter that compares 150 prediction maps of segmentation classes obtained from the generated transmission image and the ground truth image, where the pixel value in the i-th prediction map represents the relevance between the pixel location and the i-th class of semantic segmentation. We adopt the pre-trained segmentation network that uses ResNet18 [23] for an encoder and the pyramid pooling module [24] for a decoder[1]. We fix the parameters of the interpreter during training.

Fig. 5 shows the results of the semantic segmentation applied to the ground truth transmission images and input glass images, respectively. We see that the reflection artifacts on the glass images often cause inconsistent semantic context to the target transmitted scenes, and thus result in severely different segmentation maps of glass images from that of the ground truth transmission images. For example, the regions of 'seat' class on the first row and 'ceiling' class on the second row are not fully localized in glass images due to the reflection artifacts. We exploit such difference as a high-level cue for generators to remove reflection artifacts from glass images.

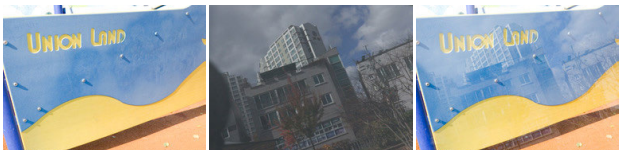
To train the proposed network, we define a loss function which consists of the color intensity loss $\mathcal{L}_{\text{col}}$ and the semantic context loss $\mathcal{L}_{\text{sem}}$. $\mathcal{L}_{\text{col}}$ is defined as the L1 norm of the difference between a predicted image and the target ground truth image, which usually makes the generators predict the pixel value of the ground truth image by referring the neighboring pixel values.

$$\mathcal{L}_{\text{col}}(\boldsymbol{\theta}) = \sum_{k=\{1,2,3\}} \left\{ ||T^k - \hat{T}^k||_1 + ||\tilde{R}^k - \hat{R}^k||_1 \right\} \quad (4)$$

where $\boldsymbol{\theta}$ denotes all parameters of the generators and k represents three image scales. T (or $\tilde{R}$) and $\hat{T}$ (or $\hat{R}$) represent the ground truth transmission (reflection) image and the generated transmission (reflection) image, respectively.

The finally generated transmission image $\hat{T}^3$ at the finest image scale is directly passed to the semantic segmentation

[1]The pre-trained model is available in https://github.com/CSAILVision/semantic-segmentation-pytorch.



(a) (b) (c)

FIGURE 6. Example of glass image. (a) Ground truth transmission image. (b) Ground truth reflection image. (c) A resulting glass image.

network $\mathcal{S}$, where we define $\mathcal{L}_{\text{sem}}$ to minimize the difference of 150 maps of predicted segmentation labels between the generated transmission image $\hat{T}^3$ and that of the ground truth transmission image T^3.

$$\mathcal{L}_{\text{sem}}(\boldsymbol{\theta}) = ||\mathcal{S}(T^3) - \mathcal{S}(\hat{T}^3)||_1. \tag{5}$$

Note that we do not consider the semantic context loss of the generated reflection images, since the reflection image usually exhibits blurred and attenuated scene contents which cause undesired semantic segmentation results. Finally, the total loss function $\mathcal{L}_{\text{total}}$ is defined as

$$\mathcal{L}_{\text{total}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{col}}(\boldsymbol{\theta}) + \kappa \mathcal{L}_{\text{sem}}(\boldsymbol{\theta}) \tag{6}$$

where κ is empirically set to 0.1.

IV. EXPERIMENTAL RESULTS

A. TEST DATA SET

We provide a new data set of real glass images including 71 triplets of glass image, transmission image and reflection image. A glass image is taken by placing a glass plane in front of camera, and the associated ground truth reflection image is captured by attaching a black opaque sheet behind the glass plane. Also, the ground truth transmission image is captured by removing the glass plane. In this paper, we show the experimental results of 23 glass images selected from the existing data sets [16]–[18] and our data set.

B. EVALUATION METHODOLOGY

The existing methods usually evaluate the performance of reflection removal in terms of the quality of the restored transmission images compared to the ground truth transmission images. However, higher scores of the transmission image quality do not always reflect better performance of reflection removal. Figs. 6 (a) and (b) show a pair of ground truth transmission and reflection images, and Fig. 6 (c) shows an associated glass image, respectively. When measuring the quantitative difference between the glass image and the ground truth transmission image, we have good scores of 0.946 in SSIM and 0.005 in LMSE, respectively, since the transmitted scene is dominant in the glass image. It means that even the lazy predictor which regards the input glass image as the restored transmission image provides good quantitative performance, which is not a desired result. On the other hand, the lazy predictor provides nothing as the reflection image, and therefore large difference is measured between the predicted reflection image and the ground truth reflection image, e.g., 0 in SSIM and 1 in LMSE.

TABLE 1. Ablation study on the performance of the proposed reflection removal algorithm.

Dataset	Evaluation		Network I	Network II	Proposed
	Metric	Image			
Zhang el al.'s test set	SSIM	T	0.700	0.718	0.755
	PSNR	T	19.45	20.55	21.11
	LMSE	T	0.053	0.048	0.042
Our test set	SSIM	T	0.788	0.783	0.824
		$\tilde{R}$	0.292	0.383	0.529
		Avg.	0.540	0.583	0.677
	PSNR	T	19.28	20.29	21.15
		$\tilde{R}$	12.99	16.81	17.94
		Avg.	16.13	18.55	19.54
	LMSE	T	0.024	0.023	0.019
		$\tilde{R}$	0.328	0.093	0.069
		Avg.	0.176	0.058	0.044

To overcome the limitation of the conventional evaluation methodology which measures the quality of the restored transmission image only, we measure the qualities of not only the restored transmission image but also the estimated reflection image compared to their ground truth images, respectively. The proposed methodology is quite simple but provides reliable evaluation results of reflection removal. Note that when existing methods do not explicitly result in reflection images, we subtract the restored transmission image from the glass image, and regard the difference image as the restored reflection image according to the linear glass image model adopted to the existing methods.

C. NETWORK TRAINING

The proposed network is trained by Adam optimizer [25] by setting the learning rate to 0.0001, and the control parameters of β_1 and β_2 to 0.5 and 0.999, respectively. In addition to our synthetic training data set, we use Zhang *et al.*'s training data set [18] together.[2] Specifically, for each epoch of training, one mini-batch is generated from Zhang *et al.*'s training data set after every 28th mini-batch generated from our training data set to balance the different sizes of the two data sets. We resize all input glass images to 256×256 and set the size of mini-batch to 12. The network is trained during 20 epochs.

D. ABLATION STUDY

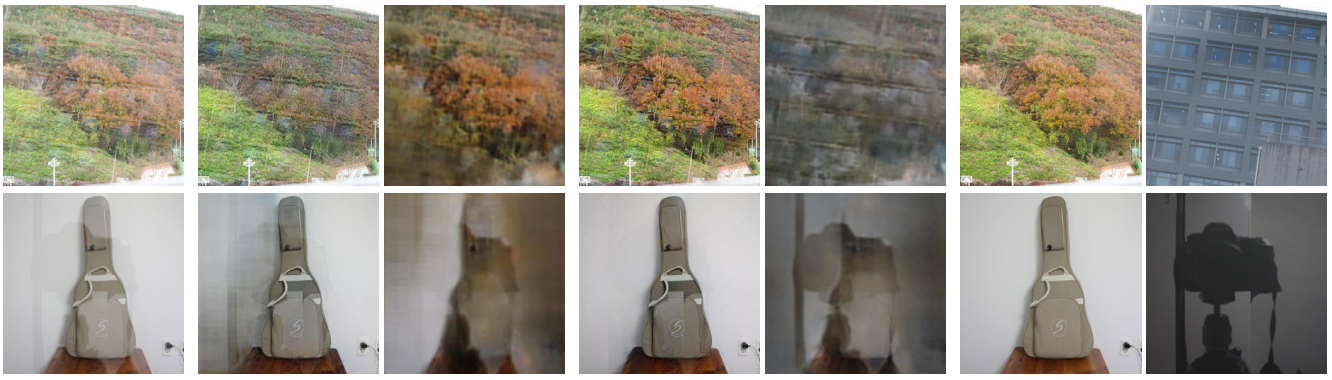
1) EFFECT OF NON-LINEARLY SYNTHESIZED TRAINING IMAGES

We first show the effect of the non-linear intensity mapping relationship used to generate synthetic training images. Table 1 compares the performance in terms of SSIM, PSNR, and LMSE. 'Network I' means the proposed network architecture which uses a base training set synthesized by an existing method [16] based on the linear glass image model.

[2]The publicly available data set provided by the authors misses one image among 90 real images for training.

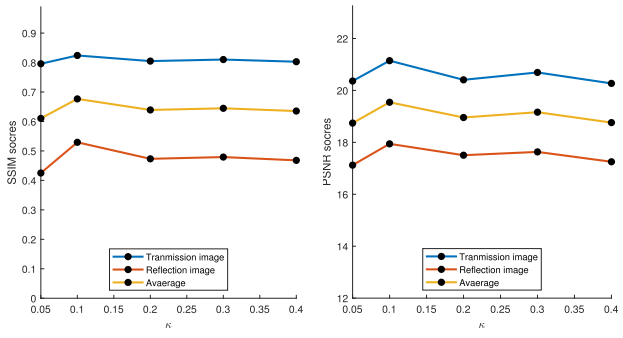
(a) (b) (c) (d)

FIGURE 7. The effect of non-linearly synthesized training images. (a) Glass images. (b) The results of Network I. (c) The results of the proposed algorithm. (d) Ground truth images. The left and right images in (b∼d) are the restored transmission and reflection images, respectively.



(a) (b) (c) (d)

FIGURE 8. The effect of semantic context. (a) Glass images. (b) The results of Network II. (c) The results of the proposed algorithm. (d) Ground truth images. The left and right images in (b∼d) are the restored transmission and reflection images, respectively.



(a) SSIM (b) PSNR

FIGURE 9. Quantitative performance of the proposed network as varying κ in Eq. (6).

We measure the qualities of the restored transmission image and the restored reflection image together on our test data set, but measure the quality of the restored transmission image only on Zhang *et al.*'s test data set since it does not provide ground truth reflection images. We see that the proposed network provides much higher quantitative performance than Network I on both of Zhang *et al.*'s test set and our test set, which demonstrates that the proposed non-linear intensity mapping relationship synthesizes more realistic glass images and thus further improves the performance of reflection removal. Also, Fig. 7 qualitatively shows the effect of using the training set synthesized by non-linear intensity mapping. Network I with the training set of the linear glass image model does not completely remove the reflection artifacts from glass images, however the proposed network using the non-linearly synthesized training set provides much better performance of reflection removal.

2) EFFECT OF SEMANTIC CONTEXT

We also investigate the effect of the semantic context loss by comparing the performance of the proposed network trained by using $\mathcal{L}_{\text{total}}$ to that of 'Network II' which is trained by using the conventional loss $\mathcal{L}_{\text{col}}$ only. As shown in Table 1, the proposed network improves the performance of Network II on both test sets in all evaluation metrics. In particular, while the semantic context loss $\mathcal{L}_{\text{sem}}$ is computed on the transmission images, the proposed algorithm improves the restoration performance on not only the transmission images but also the reflection images. Fig. 8 also shows the qualitative effect of the semantic context. While Network II without semantic loss fails to distinguish the transmission image and the reflection image completely, the proposed network yields more reliable separation results and provides quite close transmission images to the ground truth. For example,

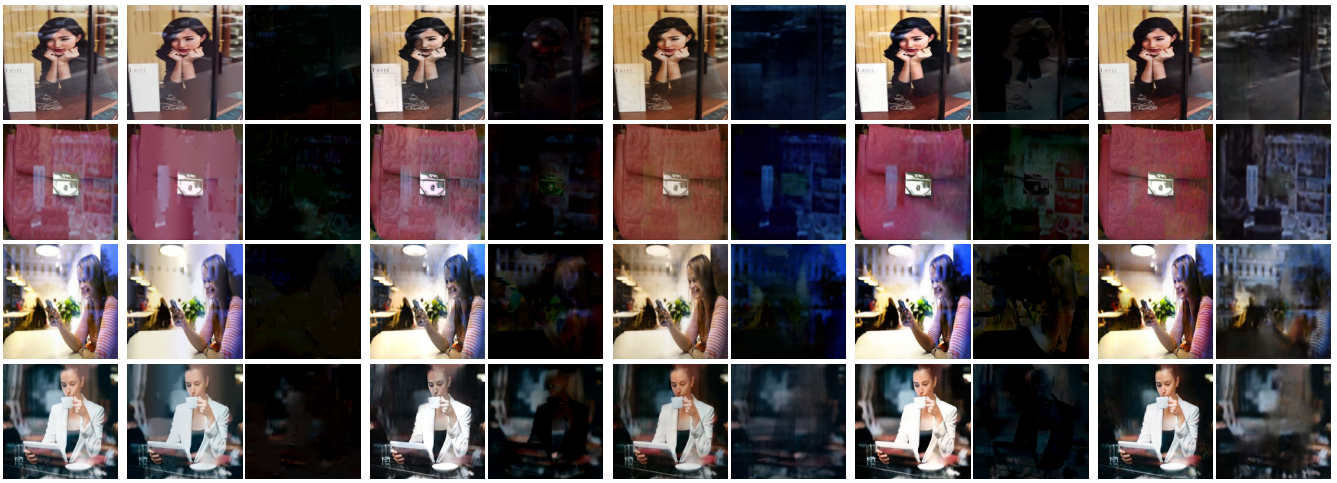


FIGURE 10. Qualitative comparison on Fan *et al.*'s test set. (a) Input glass images. The pairs of a transmission image (left) and a reflection image (right) restored by using (b) DFR [14], (c) CEILNet [16], (d) PRRNet [18], (e) BDNet [19], and (f) the proposed network.

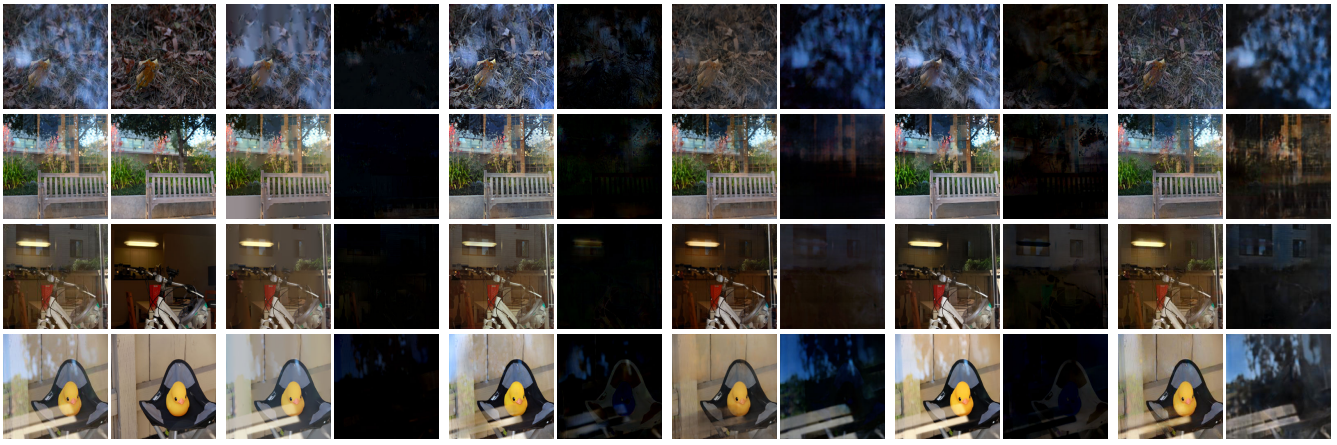


FIGURE 11. Qualitative comparison on Zhang *et al.*'s test set. (a) Input glass images. (b) The ground truth transmission images. The pairs of a transmission image (left) and a reflection image (right) restored by using (c) DFR [14], (d) CEILNet [16], (e) PRRNet [18], (f) BDNet [19], and (g) the proposed network.

the trees in the upper row and the guitar bag in the below row are restored as the transmitted scenes faithfully. For a more detailed analysis, we investigate the relative contribution of $\mathcal{L}_{\text{sem}}$ to $\mathcal{L}_{\text{col}}$ by varying κ in Eq. (6) for training the proposed network. Fig. 9 shows the associated performance on our test data set, where we see that $\kappa = 0.1$ yields the best performance in both of SSIM and PSNR.

E. PERFORMANCE COMPARISON

We compare the performance of the proposed reflection removal algorithm with that of four state-of-the-art methods: DFR [14], CEILNet [16], PRRNet [18], and BDNet [19], evaluated on real glass images in the three exiting test sets of [16]–[18] and our test set. Note that, since the existing methods do not explicitly provide restored reflection images, we generate restored reflection images by subtracting the restored transmission images from input glass images according to the linear glass image model they adopted.

1) QUALITATIVE COMPARISON

Fig. 10 first compares the results of the reflection removal on Fan *et al.*'s test set [16]. DFR distinguishes the gradients of the transmission image and the reflection image based on the degree of blurriness, but it usually fails to remove the reflection artifacts in real glass images as shown in Fig. 10(b). CEILNet adapts a deep neural network to predict the gradient maps of the transmission image for more complicated glass images, but it still remains some parts of reflection artifacts in the reconstructed transmission images as shown in Fig. 10(c). Fig. 10(d) shows that PRRNet tends to produce distorted colors in restored images, i.e., the transmission images and the reflection images tend to exhibit yellow and blue tones, respectively. BDNet employs additional networks explicitly to restore the reflection images and to generate refined transmission images, however, it also fails to provide completely separated reflection and transmission images, as shown in Fig. 10(e). In contrary, the proposed algorithm

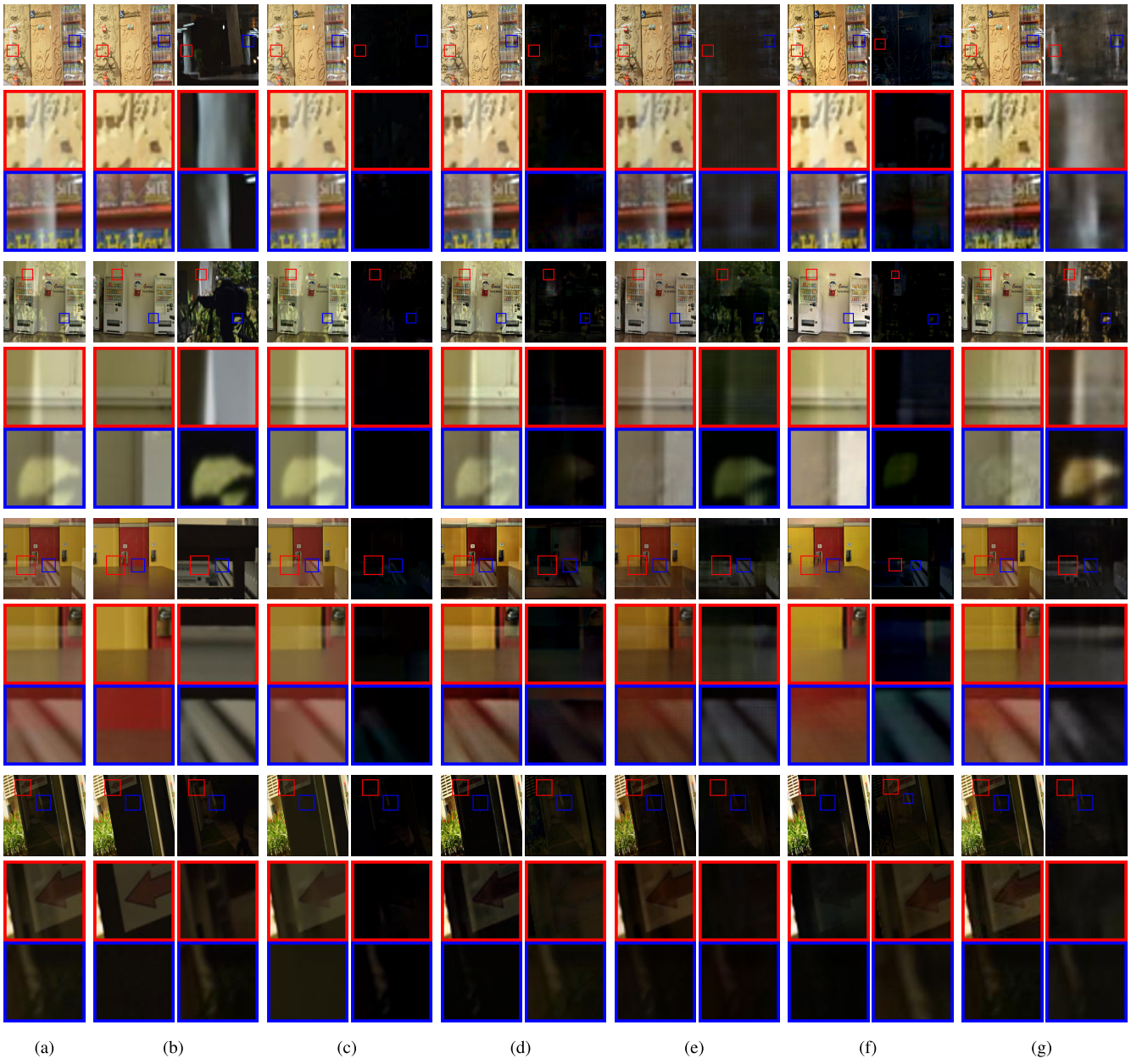


FIGURE 12. Qualitative comparison on Wan *et al.*'s test set. (a) Input glass images. (b) The ground truth transmission (left) and reflection (right) images. The pairs of a transmission image (left) and a reflection image (right) restored by using (c) DFR [14], (d) CEILNet [16], (e) PRRNet [18], (f) BDNet [19], and (g) the proposed network.

restores both of the transmission and reflection images more faithfully compared with the existing methods. Especially, the proposed algorithm outperforms the existing methods in terms of the quality of the restored reflection images, for example, the fence wall is well restored in the reflection image of the proposed algorithm as shown in the third row.

Fig. 11 shows the qualitative comparison results on Zhang *et al.*'s test set [18] where the ground truth transmission images are available. We see similar results to Fig. 10 since the proposed algorithm outperforms DFR, CEILNet, and BDNet in most cases. However PRRNet sometimes yields relatively better performance than the proposed algorithm, e.g., the forth test image in Fig. 11.

Fig. 12 also provides the comparison results on Wan *et al.*'s test set [17] where both of the transmission and reflection images have their ground truth. The red and blue boxes highlight local image regions where the reflection artifacts are clearly observed. BDNet yields relatively good performance on Wan *et al.*'s test set compared to Fan *et al.*'s and Zhang *et al.*'s, e.g., the third test image on which the other methods usually fail to work. However, it often removes strong visual structures from the resulting transmission images, e.g., the floor and the arrow sign are disappeared

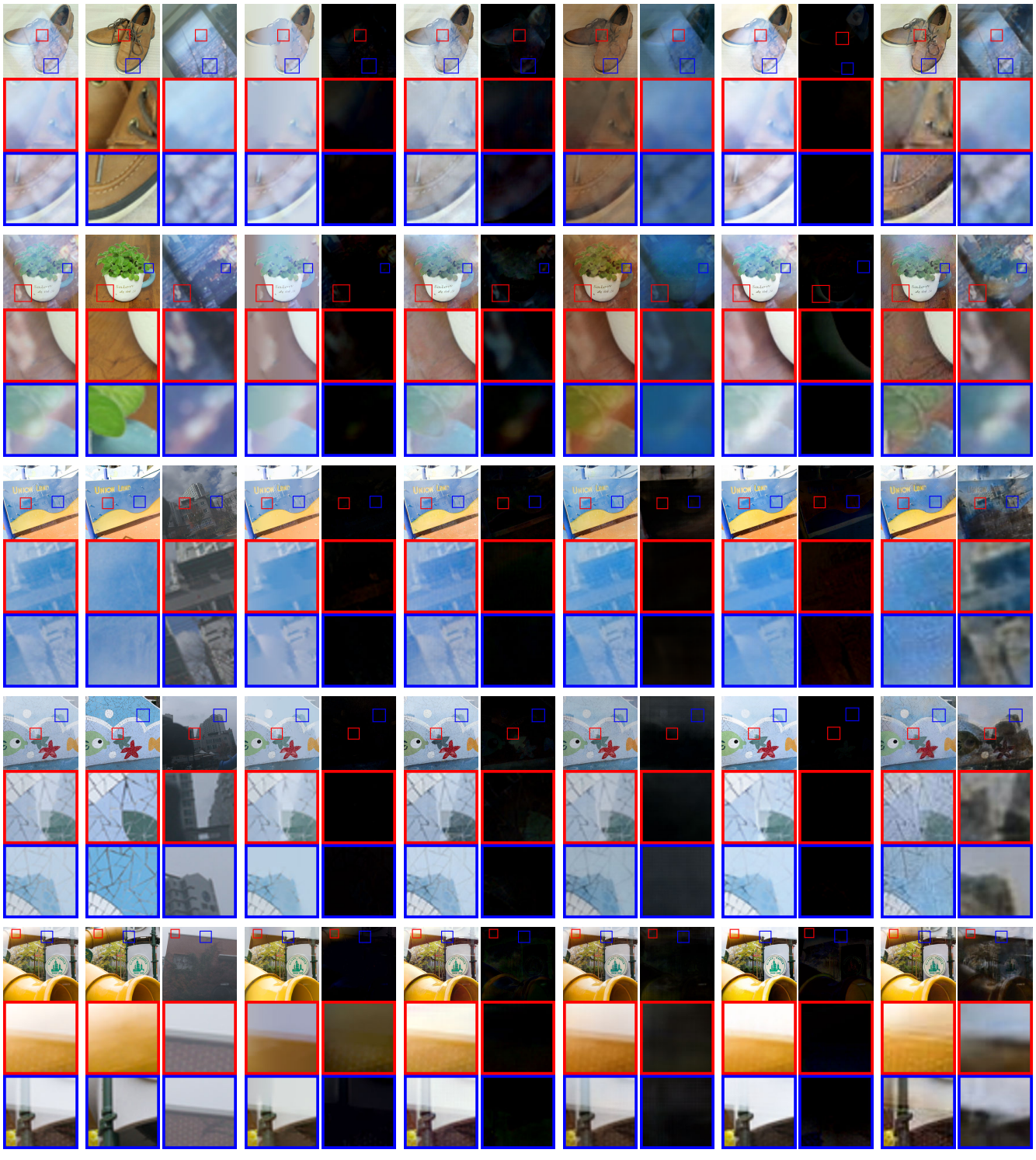


FIGURE 13. Qualitative comparison on our test set. (a) Input glass images. (b) The ground truth transmission (left) and reflection (right) images. The pairs of a transmission image (left) and a reflection image (right) restored by using (c) DFR [14], (d) CEILNet [16], (e) PRRNet [18], (f) BDNet [19], and (g) the proposed network.

in the third and fourth test images. Note that the proposed algorithm provides better or comparable performance compared to BDNet, and always outperforms DFR and CEILNet.

In addition to the existing test sets, Fig. 13 compares the reflection removal results on our test set which provides the ground truth of both of the transmission and reflection images. As shown in Figs. 13 (c-f), the existing methods

TABLE 2. Comparison of quantitative performance of reflection removal. The best and the second best scores are denoted in red and blue, respectively.

Test DB	Metric	Image	DFR	CEILNet	PRRNet	BDNet	Prop.
Zhang *et al.*'s test set	SSIM	T	0.659	0.704	0.805	0.672	0.755
	PSNR	T	18.15	18.55	22.20	18.52	21.11
	LMSE	T	0.055	0.053	0.035	0.063	0.043
Wan *et al.*'s test set	SSIM	T	0.882	0.842	0.857	0.828	0.908
		$\tilde{R}$	0.244	0.337	0.488	0.350	0.615
		Avg.	0.563	0.590	0.673	0.589	0.762
	PSNR	T	23.66	22.72	22.47	22.06	25.48
		$\tilde{R}$	20.33	20.29	21.67	20.91	22.99
		Avg.	21.99	21.51	22.07	21.49	24.24
	LMSE	T	0.013	0.023	0.021	0.020	0.012
		$\tilde{R}$	0.606	0.558	0.237	0.534	0.160
		Avg.	0.309	0.291	0.129	0.277	0.086
Our test set	SSIM	T	0.768	0.801	0.739	0.787	0.824
		$\tilde{R}$	0.068	0.092	0.400	0.088	0.529
		Avg.	0.418	0.447	0.569	0.438	0.677
	PSNR	T	18.12	19.37	18.03	18.13	21.15
		$\tilde{R}$	10.81	10.98	15.05	11.08	17.94
		Avg.	14.46	15.17	16.54	14.60	19.54
	LMSE	T	0.027	0.023	0.023	0.022	0.020
		$\tilde{R}$	0.592	0.611	0.119	0.678	0.060
		Avg.	0.309	0.317	0.071	0.351	0.040

fail to recognize the reflection artifacts in glass images, and hence result in mostly black colors in the restored reflection images. Accordingly, the resulting transmission images are almost similar to the input glass images. On the contrary, as shown in Fig. 13 (g), the proposed algorithm restores both of the ground truth transmission and reflection images more faithfully and achieves much better performance of reflection removal compared with the existing methods.

2) QUANTITATIVE COMPARISON

Table 2 shows the quantitative performance of the proposed algorithm compared with that of the existing methods in terms of the three metrics of SSIM, PSNR, and LMSE. We use 20 glass images of Zhang *et al.*'s test set, 55 glass images of Wan *et al.*'s test set, and 71 glass images of our test set. We measure the qualities of the restored transmission and reflection images together compared to their ground truth images, however, only the transmission images are evaluated on Zhang *et al.*'s test set due to the lack of ground truth reflection images. Note that DFR, that rarely removes reflection artifacts in qualitative comparison results, provides relatively better scores with respect to the transmission images, but it yields bad scores with respect to the reflection images on Wan *et al.*'s test set and our test set. It means that our quantitative evaluation methodology of considering both of the transmission and reflection images together provides more reliable results. For Zhang *et al.*'s test set, PRRNet provides the best scores in term of all the three evaluation metrics, whereas the proposed algorithm achieves the second best scores. However, for Wan *et al.*'s test set and our test set, the proposed algorithm yields the best scores in all metrics.

(a) (b) (c)

FIGURE 14. Limitation of the proposed algorithm. (a) Glass images. (b) The results of the proposed algorithm. (c) Ground truth images. The left and right images in (b) and (c) are the restored transmission and reflection images, respectively.

V. CONCLUSION

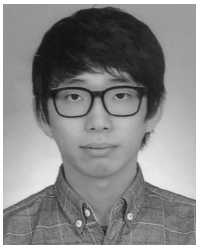
In this paper, we proposed a single image reflection removal algorithm. We first suggested a non-linear intensity mapping for glass images to synthesize more realistic training sets. We designed a network using multi-scale generators and an interpreter which employs the semantic context of the transmission images for reflection removal. We also generated a new test data set of 71 real glass images with ground truth transmission and reflection images. Experimental results demonstrated that the proposed algorithm restores both of the transmission and reflection images more faithfully compared with the existing methods.

Future research includes the design of fully automatic deep neural networks which employ the characteristics of non-linear glass image model. It is expected to overcome the limitation of the proposed algorithm especially tested on challenging glass images as shown in Fig. 14, where intensive reflection artifacts still remain in the restored transmission images.

REFERENCES

[1] L.-Q. Zuo, H.-M. Sun, Q.-C. Mao, R. Qi, and R.-S. Jia, "Natural scene text recognition based on encoder-decoder framework," *IEEE Access*, vol. 7, pp. 62616–62623, 2019.
[2] C. Yao, P. Sun, R. Zhi, and Y. Shen, "Learning coexistence discriminative features for multi-class object detection," *IEEE Access*, vol. 6, pp. 37676–37684, 2018.
[3] Z. Yi, T. Chang, S. Li, R. Liu, J. Zhang, and A. Hao, "Scene-aware deep networks for semantic segmentation of images," *IEEE Access*, vol. 7, pp. 69184–69193, 2019.
[4] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proc. IEEE CVPR*, Jun. 2014, pp. 2752–2759.
[5] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proc. IEEE CVPR*, Jun. 2015, pp. 3193–3201.
[6] X. Guo, X. Cao, and Y. Ma, "Robust separation of reflection from multiple images," in *Proc. IEEE CVPR*, Jun. 2014, pp. 2187–2194.
[7] Y. Y. Schechner, N. Kiryati, and J. Shamir, "Blind recovery of transparent and semireflected scenes," in *Proc. IEEE CVPR*, Jun. 2000, pp. 38–43.
[8] A. Agrawal, R. Raskar, and R. Chellappa, "Edge suppression by gradient field transformation using cross-projection tensors," in *Proc. IEEE CVPR*, Jun. 2006, pp. 2301–2308.

[9] A. Agrawal, R. Raskar, S. K. Nayar, and Y. Li, "Removing photography artifacts using gradient projection and flash-exposure sampling," in *Proc. ACM SIGGRAPH*, Jul. 2005, pp. 828–835.
[10] N. Kong, Y.-W. Tai, and J. S. Shin, "A physically-based approach to reflection separation: From physical modeling to constrained optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 209–221, Feb. 2014.
[11] S. N. Sinha, J. Kopf, M. Goesele, D. Scharstein, and R. Szeliski, "Image-based rendering for scenes with reflections," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 100:1–100:10, Jul. 2012.
[12] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 79:1–79:11, 2015.
[13] B.-J. Han and J.-Y. Sim, "Glass reflection removal using co-saliency-based image alignment and low-rank matrix completion in gradient domain," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4873–4888, Oct. 2018.
[14] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot, "Depth of field guided reflection removal," in *Proc. IEEE ICIP*, Sep. 2016, pp. 21–25.
[15] A. Levin, A. Zomet, and Y. Weiss, "Separating reflections from a single image using local features," in *Proc. IEEE CVPR*, Jun./Jul. 2004, p. 1.
[16] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE CVPR*, Oct. 2017, pp. 3238–3247.
[17] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "CRRN: Multi-scale guided concurrent reflection removal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4777–4785.
[18] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. IEEE CVPR*, Jun. 2018, pp. 4786–4794.
[19] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proc. ECCV*, Jun. 2018, pp. 654–669.
[20] *Image Process. Toolbox: Reference*, MATLAB, MathWorks, Natick, MA, USA, 2018.
[21] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE CVPR*, Jul. 2017, pp. 5122–5130.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770–778.
[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE CVPR*, Jul. 2017, pp. 6230–6239.
[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015, pp. 1–15.



BYEONG-JU HAN received the B.S. degree in electrical and computer engineering from the Ulsan National Institute of Science and Technology, Ulsan, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include computer vision and machine learning.

JAE-YOUNG SIM (S'02–M'06) received the B.S. degree in electrical engineering, and the M.S. and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 1999, 2001, and 2005, respectively.

From 2005 to 2009, he was a Research Staff Member of the Samsung Advanced Institute of Technology, Samsung Electronics Company, Ltd. In 2009, he joined the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea, where he is currently an Associate Professor. His research interests include image, video, and 3D visual processing, computer vision, and multimedia data compression.