Intensity-aware GAN for Single Image Reflection Removal

Nien-Hsin Chou, Li-Chung Chuang, and Ming-Sui Lee
Graduate Institute of Networking and Multimedia
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan

E-mail: {r03944029, d02944014, mslee}@csie.ntu.edu.tw Tel: +886-2-33664888

***Abstract*— Single image reflection removal is a challenging task in computer vision. Most existing approaches rely on carefully handcrafted priors to solve the problem. Contrast to the optimization-based methods, an intensity-aware GAN with dual generators is proposed to directly estimate the function which transforms the mixture image into the reflection image itself. From the observation that the reflection layer has more discriminating power in the region with low intensity than that in the region with high intensity, the proposed architecture better describes the characteristic of the model. Moreover, a reflection image synthesis method based on the screen blending model is also presented. Experimental results demonstrate that the results of reflection removal are satisfactory in real cases while comparing with state-of-the-art methods.**

I. INTRODUCTION

Images of a scene taken through transparent or translucent material like glasses are often plagued by undesired artifacts such as the reflected hallucinations of the foreground scene. As shown in Fig. 1, the captured image I is superimposed by the reflection image R and the intended reflection-free image which is called the transmission image T. We attempt to separate the transmission image from the reflection image since it not only improves the visual quality but also benefits many computer vision tasks such as object detection or image classification. This separation is intuitive for human beings but is very challenging for computer vision algorithms.

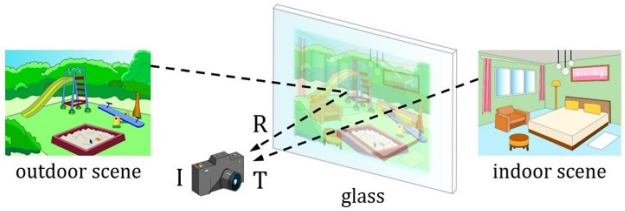


Fig. 1: The mixture image of the reflection and the transmission.

Such a problem is ill-posed due to the infinite combinations of the transmission image and the reflection image. Several research works tackled this problem by exploiting redundant information with multiple images [1,2,3,4]. However, multiple images are not always available and it often requires precise image registration so that it is not applicable to moving objects such as taking photo through the window on a train. Other works taking single input image tried to impose additional constraints using prior knowledge such as the statistics of the natural images or the guidance of user inputs [5,6,7,8,9]. But it is known that methods which need user input are inconvenient on practice. As the optimization-based methods focus on finding proper priors to model the mixture image, the rise of deep neural network [10] provides an alternative thought to this problem [11][12]. One may directly learn the function itself from the data in an end-to-end manner, and then infer the answer via the function.

Here we claim that the reflection removal problem can be viewed as an image-to-image translation problem: given a mixture image I, find a function $f(x; \theta)$ which translates I to its reflection-free transmission image T. The image-to-image framework [13] which is based on the Generative Adversarial Network (GAN) [14] is adopted as the base network. The image-to-image network trains a GAN in two parts: a generator network to translate a source image into a target image, and a discriminator network to distinguish whether the generated image along with the source image form a pair or not. In other words, we translate a mixture image into a reflection image.

A well-known issue of deep neural network is that it relies on large amount of data. However, there's no such particular dataset existing for reflection removal. In our task, an example consists of a reflection image and a "clean" image. The process of obtaining both images from real scene requires special hardware and costs a lot of time. Consider the total amount of data we need, synthesizing the data from real images is a more tractable way. In contrast to the commonly adopted additive blending model, a screen blending model is chosen to mimic the intertwining effect between the reflection image and the transmission image. From the observation that the reflection is more prominent in the low intensity region than that in the high intensity region, a model with dual generators is proposed to describe the characteristics of the low intensity region and high intensity region separately. Comparing to other image-to-image framework where L1-loss or L2-loss is utilized to regularize the generated image as close as possible to the input image, an L1-loss weighted by the reflection intensity is proposed to model the reflection phenomenon adaptively.

The contributions of the proposed work are listed below. 1) A screen blending model is proposed to synthesize the reflection images based on a multiplicative model instead of the additive one. The experimental results show that the

proposed model outperforms the traditional one in terms of the trueness of the real phenomenon. 2) A generative adversarial network with dual generators is proposed to compromise the different characteristics between the high intensity regions and low intensity regions. 3) A loss function based on the weighted L1-norm provides more adaptability than the original image-to-image framework.

The paper is organized as follows. An overview of the related work of reflection removal is presented in Section II and the proposed method is detailed in Section III. Section IV demonstrates the experimental results and the conclusive remarks are given in Section V.

II. Related Work

The related work of reflection removal can be categorized into two groups based on the type of input. Each of them is reviewed in the following two subsections.

A. Multiple Images Reflection Removal

Lots of work rely on multiple input images to find cues which help to distinguish the transmission layer from the reflection layer. Li and Brown [1] proposed to exploit a common phenomenon that the reflection varies with respect to the background when viewing angle changes. By aligning the images taken from different viewpoints, gradients with variation are likely to belong to the reflection and gradients belong to the background are mostly remain constant. Sun et al. [2] leverages motion cue to separate the reflection and the transmission layer. They observed that the background pixels tend to dominate the motion vectors across different SIFT-flow smoothness levels, and pixels from the reflection layer are less prominent than the background. Guo et al. [3] used three priors to find optimized separation: correlation of transmission layer in images, the sparsity and the independence between the gradient fields of the transmission and the reflection layer. Xue et al. [4] used motion difference to find the initial transmission and reflection layers. An iterative procedure based on dense motion fields is conducted to refine the transmission layer and reflection layer. Although methods with multiple input images achieve remarkable results, the requirement of multiple inputs hinder its usage in practical situations.

B. Single Image Reflection Removal

Single image based methods rely on priors to constrain the solution space. Levin and Weiss [5] proposed to leverage the sparse property of natural image by imposing a Laplacian prior over the gradients. Two sets of points with user labeling are required to regularize the optimization. Li and Brown [6] proposed to characterize the transmission layer and the reflection layer by their gradient histogram distribution as the transmission layer tends to have large gradients and the reflection layer is often blurred and smooth. Shih et al. [7] observed that most reflection images come with "ghosting" artifacts and can be modeled as the convolution with an estimated kernel. To further ease the ill-posed problem, patch based GMM priors are imposed on both the transmission and the reflection layer. Wan et al. [8] use DoF (depth of field) as the main clue to distinguish the edges belonging to the background from those belong to the reflection layer. With the estimated reflection edge map and the transmission edge map, Levin and Weiss's method [5] attempts to reconstruct both the transmission and reflection images. Based on a Laplacian data fidelity term and the gradient sparsity term imposed on the output, Arvanitopoulos et al. [9] tries to suppress the reflection effect rather than removing it completely.

The proposed method differs from the above optimization based methods which solves for the answers with certain constraints. We directly estimate the transform function from I to T via a deep neural network.

III. The Proposed Method

A. Data Synthesis

An unavoidable problem comes along with the deep learning is that it requires massive amount of training data. In the case of reflection removal, each sample is an aligned pair of a source mixture image and a target transmission image. Although it is possible to acquire both images through specialized optical device such as polarizing filters, the process is time-consuming and inefficient. Thus, we opt for using synthetic data as our training samples. The filtered Open Images dataset [15] is considered as our natural image source. It consists over 9 million images with 6000 categories which provides sufficient diversity of data. First, two images are randomly selected from the dataset. One of them is conducted with intensity adjustment ranging from 9% to 18% stochastically, which results as the reflection layer. The reason why this range of the percentage is chosen can be explained via Fig. 2. The reflectance of glass usually falls between 0.05 and 0.1. Assume the energy of the incident light is E and the reflectance of the given medium is 0.1. In this case, the first reflection R1 is 0.1E and the second reflection R2 is 0.081E. The energy of third reflection R3 is less than 0.001E which can be neglected. As a result, the total energy of the reflection captured by the camera is about 0.181E (18.1%). Similarly, the reflection is about 0.095E (9.5%) when the reflectance of the medium is 0.05.

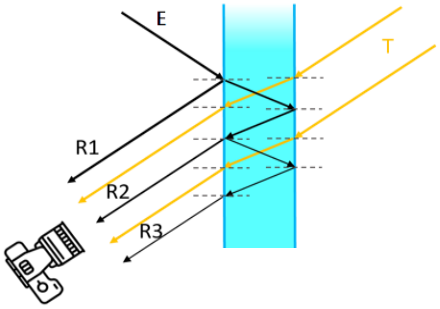


Fig. 2: The illustration of energy decay of the reflection light.

The mapping function is regressed from real images captured with different shutter time and Fig. 3 shows the regression of data clusters. The other one, as the transmission layer, remains unchanged in the final compositing process.

The following equation describes a widely adopted additive model.

$$I = T + R \tag{1}$$

Even though it is satisfactory in most image compositing cases, it is too simplified to model the complicated intertwining factors between the transmission image and the reflection one.

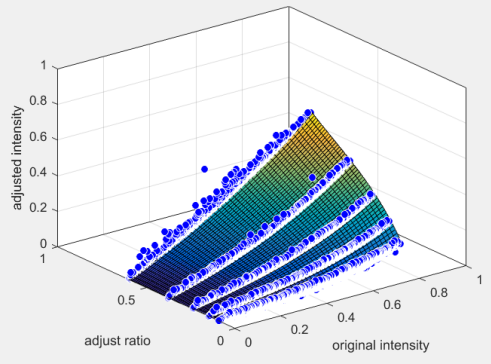


Fig. 3: Surface regression of intensity with different energy ratios.

Therefore, we adopt the screen model for image blending:

$$\mathrm{I} = 1 - (1 - \mathrm{T}) \odot (1 - \mathrm{R}) = T + (1 - T) \odot R, \quad (2)$$

where $\odot$ denotes the Hadamard product. This essentially equals to the additive model where the reflection image is weighted by the inverse intensity of the transmission image, which is in accordance with our observation of realistic mixture images. In order to verify the screen model, an experiment is performed as follows. First, the transmission layer, T, is taken without any obstacles in between the camera and the indoor scene. Next, a piece of glass is set between the camera and the indoor scene in order to capture the mixture image which contain the transmission layer and the reflection layer. The last step, as illustrated in Fig. 4, a black curtain is attached to the glass while shooting the scene so that the real reflection layer, R, (the outdoor scene) can be obtained purely. Once the real transmission image and real reflection image are available, the image synthesis process is conducted according to the additive model and the screen model separately.

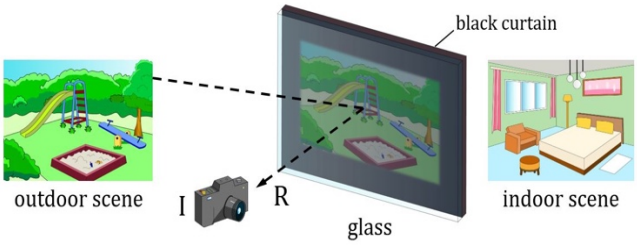
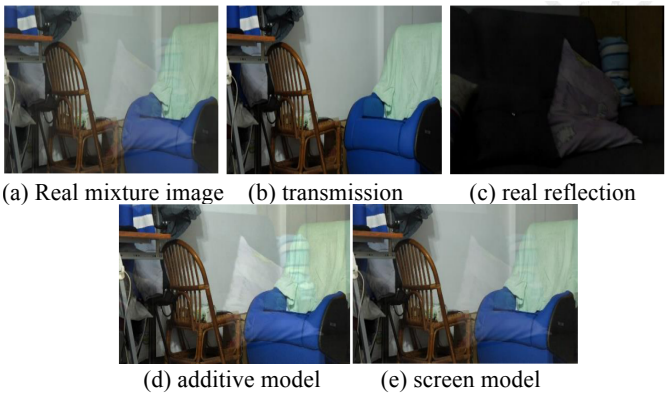


Fig. 4: The way to capture the real transmission layer, R.

Fig. 5 demonstrates the difference between the additive blending model and the screen blending model. As can be seen from those figures, the result (Fig. 5(e)) generated by the screen model is much more similar to the real mixture image (Fig. 5(a)), which explains the reason why screen model is adopted in synthesizing the training data. More synthetic images are shown in Fig. 6.



(a) Real mixture image (b) transmission (c) real reflection

(d) additive model (e) screen model

Fig. 5: Comparison of blending results.

Fig. 6: Several synthetic images.

B. *Intensity-aware Single Image Reflection Removal*

Isola et al. [13] proposed a new framework based on conditional Generative Adversarial Network [14] to solve the image-to-image translating problem. GANs formulate the network learning as a competitive game between a generator G and a discriminator D. The generator G attempts to generate fake samples as close as possible to the real samples from the dataset, and the discriminator is trained to differentiate the fake samples. The objective function of GAN is

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}\_(x, z)\ [\log\ (1 - D(G(x, z)))\ ]. \quad (3)$$

In the image-to-image setting, the discriminator is conditional on the input image. Thus, the objective function became

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))]. \quad (4)$$

To further encourage the output of the generator to resemble the input ground truth image, an L1-loss or L2-loss is added empirically.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1] \quad (5)$$

In order to comply with the screen blending model assumption, the L1-loss between the generated reflection image and the ground truth reflection image is weighted by the inverse of the ground truth transmission image, which results in emphasizing the importance on the low intensity region. This modified L1-loss is

$$\mathcal{L}_{WL1\_low}(G) = \mathbb{E}_{x,y,z}[\|(1 - T_{gt}) \odot (R_{gt} - G(x,z))\|_1]. \quad (6)$$

For better modeling the characteristic difference between the high intensity region and the low intensity region, a GAN with dual generators G_1 and G_2 is proposed and is shown in Fig. 7.

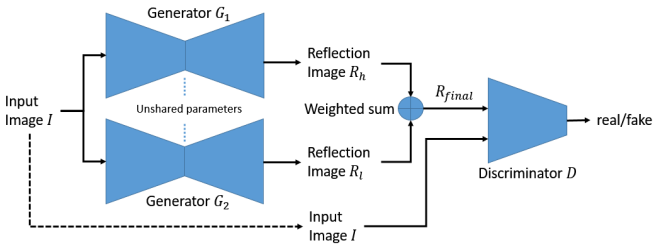


Fig. 7: The proposed GAN with dual generators.

Generator G_1 is responsible for generating the high intensity region of the reflection image, while generator G_2 is for low intensity region. The L1-loss for G_1 is simply weighted by the ground truth transmission intensity T_{gt}, which is

$$\mathcal{L}_{WL1\_high}(G) = \mathbb{E}_{x,y,z}[\|T_{gt} \odot (R_{gt} - G(x,z))\|_1], \quad (7)$$

and the overall objective is

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \mathcal{L}_{WL1_{low}}(G) + \mathcal{L}_{WL1\_high}(G). \quad (8)$$

Therefore, the final reflection image can be obtained as follows:

$$R_{final} = (1 - T_l) \odot R_l + T_l \odot R_h \quad (9)$$

where T_l, R_l is the transmission layer and the reflection layer generated by the low region generator, R_h is the reflection layer generated by high region generator and R_{final} is the final generating result and the input of discriminator.

IV. Experimental Results

For the generator G, the U-net, which is an auto encoder with symmetric skip connections is adopted. For the discriminator D, the PatchGAN is utilized as in the original image-to-image network. 100K color images with resolution 256×256 is synthesized for the training set and several real photos with reflection are gathered for testing purpose. The U-Net is pre-trained with the L1-loss for 50 epochs, which provides a nice initialization for the generator to start with and also helps the GAN training. The full GAN network is trained for 100 epochs afterwards.

A. Qualitative Results

Fig. 8 shows several results of the proposed method. Fig. 8(a) is the input image, Fig. 8(b) and Fig. 8(c) are the reflection layers of low intensity region and high intensity region, respectively. Fig. 8(d) represents the composited reflection layer and Fig. 8(e) shows the transmission layer extracted by the proposed method. As we can see, the reflection layers and the transmission layer are satisfactory.

B. Comparison with state-of-the-art methods

In this section, three single-image-based approaches are compared: Li and Brown [6], Wan et al. [8] and Arvanitopoulos et al. [9] as state-of-the-art. Fig. 9 demonstrates the comparing results with reflection strength increasing from top to bottom. As can be seen from the results of first three sets, Li and Brown [6] removes the reflection partially at the cost of causing obvious color deviation. Wan et al. [8] has better capability of eliminating reflection but lots of texture details are wiped out as well. Arvanitopoulos et al. [9] seems to over smooth the image in the regions which contain reflection, which results in undesired artifacts. Overall speaking, the results of the proposed method not only keep the similar color tone with the input image but also remain the details of the texture parts. The most importance thing is that the reflection layer can be removed successfully. For the last set of the results whose input image has the strongest power of reflection, none of the methods perform satisfactorily. Even though the result of the proposed method still retains the detail and color tone of the input image, the reflection can only be cleaned to a limited degree. In this case, the removal process may be executed iteratively to obtain a better result.

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

Fig. 8: (a) input I, (b)(c)(d) reflection layers, (e) transmission T.

V. Conclusion

In this paper, we tackle the single image reflection removal problem by casting it as an image-to-image translation task, which is benefited from the rich modeling ability of deep neural network. A GAN with dual generators is proposed to model the reflection regions with different intensity separately. The training image are synthesized based on the screen blending model as we claim that it is more accurate to describe the real phenomenon than the commonly used additive model. The experimental results demonstrate that the proposed method provides competitive results of reflection removal. However, the cases with strong reflection power still remain challenging.

References

[1] Y. Li and M. S. Brown, "Exploiting Reflection Change for Automatic Reflection Removal," in *IEEE International Conference on Computer Vision (ICCV)*, pages 2432–2439, 2013.

(a) Input image (b) [6] (c) [8] (d) [9] (e) proposed method

Fig. 9: Performance comparison with other methods.

[2] C. Sun, S. Liu, T. Yang, B. Zeng, Z. Wang, and G. Liu, "Automatic Reflection Removal using Gradient Intensity and Motion Cues," in *Proceedings of the ACM on Multimedia Conference*, 2016.

[3] X. Guo, X. Cao, and Y. Ma, "Robust Separation of Reflection from Multiple Images," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2195–2202, 2014.

[4] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A Computational Approach for Obstruction-free Photography," *ACM Transactions on Graphics*, 34(4):79:1–79:11, 2015.

[5] A. Levin and Y. Weiss, "User Assisted Separation of Reflections from a Single Image Using a Sparsity Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI),* 29(9):1647–1654, 2007.

[6] Y. Li and M. S. Brown, "Single Image Layer Separation Using Relative Smoothness," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection Removal using Ghosting Cues," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3193–3201, 2015.

[8] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot, "Depth of Field Guided Reflection Removal," in *IEEE International Conference on Image Processing (ICIP)*, pp. 21–25, 2016.

[9] N. Arvanitopoulos, R. Achanta, and S. Süsstrunk, "Single Image Reflection Suppression," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," in *Nature* 521.7553 (2015): 436.

[11] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A Generic Deep Architecture for Single Image Reflection Removal and Image Smoothing," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] X. Zhang, R. Ng, and Q. Chen, "Single Image Reflection Separation with Perceptual Losses," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Neural Information Processing Systems (NIPS)*, 2014.

[15] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, V. Gomes, A. Gupta, D. Narayanan, C. Sun, G. Chechik, and K. Murphy, "OpenImages: A public dataset for large-scale multi-label and multi-class image classification," (https://github.com/openimages), 2016.