The following publication X. Wang, Y. Pan and D. P. K. Lun, "Stereoscopic image reflection removal based on Wasserstein Generative Adversarial Network," 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), 2020, pp. 148-151 is available at https://dx.doi.org/10.1109/VCIP49819.2020.9301892.

Stereoscopic image reflection removal based on Wasserstein Generative Adversarial Network

Xiuyuan Wang, Yikun Pan and Daniel P.K. Lun*

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong
enpklun@polyu.edu.hk

***Abstract*—Reflection removal is a long-standing problem in computer vision. In this paper, we consider the reflection removal problem for stereoscopic images. By exploiting the depth information of stereoscopic images, a new background edge estimation algorithm based on the Wasserstein Generative Adversarial Network (WGAN) is proposed to distinguish the edges of the background image from the reflection. The background edges are then used to reconstruct the background image. We compare the proposed approach with the state-of-the-art reflection removal methods. Results show that the proposed approach can outperform the traditional single-image based methods and is comparable to the multiple-image based approach while having a much simpler imaging hardware requirement.**

Keywords—Reflection removal, GAN, stereoscopic images

I. INTRODUCTION

Images captured through a semi-transparent material, such as glass, are often spoiled by a superimposed layer of reflection image. The intended transmission layer, which is regarded as the background image, is intertwined with the reflection part, causing annoying noises in the image and thus making various image restoration tasks difficult, if not impossible [1]. Mathematically, the degradation model is given as follows:

$$I = I_R + I_B, \tag{1}$$

where I_R and I_B are the reflection and background respectively. The task of reflection removal is to restore I_B from I. The problem is ill-posed as we have more unknowns than observation. To reduce the ill-posedness, conventional methods introduce different priors into their algorithms based on the features of the reflection and background images [2-5]. However, due to the huge variation of natural images, these priors can hardly be valid for all images captured in different conditions. It affects the robustness of these approaches. Besides, these methods are usually time-consuming due to the massive optimization processes in their algorithms. As human beings are good at distinguishing the background from the reflection, many learning based approaches were developed recently. In particular, various deep learning methods were researched for solving the reflection removal problem [1,6-10]. Current deep learning based reflection removal methods can be classified as single or multiple-image based. Comparing with the multiple-image ones, the single-image based methods are more general. However, their limitations are obvious due to the challenging nature of the problem. To lower the difficulty, most of these approaches assume the reflection is blurry, which is however often not the case in practice. The multiple-image based approaches allow the use of the depth cue to distinguish the background and reflection. It is a relatively general prior as it is seldom to have two uncorrelated scenes (background and reflection) having the same depth range. However, it is never trivial to obtain the depth of an image with reflection since in this case every pixel will have two depths. It is referred as the depth ambiguity problem [6]. To solve the problem, [6] uses array images with 5 views to estimate the depth. It however lowers the generality of the method as it is uncommon to have array images with 5 views in daily photography.

With the advance in imaging technology, current mobile devices are often equipped with stereoscopic cameras. It allows stereoscopic images to be more readily available than in the past. By having stereoscopic images, it is possible to acquire the depth information in an image for solving the reflection removal problem. In this paper, we propose a reflection removal approach for stereoscopic images. To avoid the depth ambiguity problem as mentioned above, we focus only on the strong gradients of the images. It is based on the edge independent property that two uncorrelated images (background and reflection) seldom have their edges overlapped [6]. We propose a stereoscopic background edge estimation algorithm to estimate the background edges directly from the input stereoscopic images. The edges are then fed to an image reconstruction network [6] to obtain the background image. For the proposed algorithm, a Block Matching (BM) approach is used to first estimate the disparities along the image edges. Then we classify the edges as background or reflection based on their disparities. Due to the possible classification errors, we keep only the image edges that we are confident of their class. The rest is removed and regenerated by a Wasserstein Generative Adversarial Network (WGAN) [11]. We compared the proposed method with the 5-view based approach [6] and two other single-image based approaches [7-8]. Our experimental results show that the proposed approach has a much better performance than the single-image based methods in removing strong reflections. It can also achieve a

This work is fully supported by the Hong Kong Polytechnic University under the research grant ZZJV.

competitive result compared with the 5-view based approach while requiring much simple hardware to acquire the images.

II. The Proposed Algorithm

The operation flow of the proposed stereoscopic background edge estimation algorithm is shown in Fig. 1. As mentioned in Section I, the depth ambiguity problem introduces much difficulty when estimating the disparity of an image with reflection. Based on the edge independent property [6], we deal with the depth ambiguity problem by evaluating only the disparities along the strong gradients. It is because the background and reflection images, which are usually uncorrelated, are seldom to have their edges overlapped. As shown in Fig. 1, the gradient magnitudes are firstly extracted from the input stereoscopic images. We introduce a thresholding process such that only the strong gradients are retained. The threshold was set to 0.2 for a normalized image.

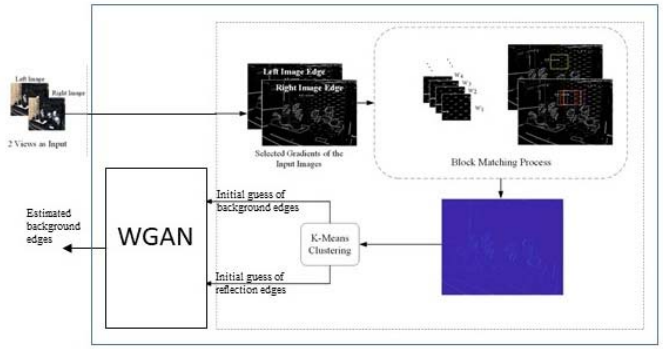


Figure 1. Operation flow of the proposed algorithm

Then, a BM approach is used for estimating the disparities along the strong gradients. Define the strong gradient maps of the left and right stereoscopic images as E_{left} and E_{right}, respectively. The BM operation can be described by (2).

$$d^*(x,y) = \underset{d\in(0,D_{max}]}{\arg\min} \sum_{i=-W/2}^{W/2} \sum_{j=-H/2}^{H/2} \left\| E_{right}\,(x+i+d, y+j) - E_{left}(x+i, y+j) \right\|_2^2, \quad (2)$$

where (W, H) is the width and height of the block and $d^*(x,y)$ is the estimated disparity at a strong gradient pixel (x,y) of the left image. Collecting the disparity $d^*(x,y)$ for all (x,y) forms the edge disparity map. Fig. 2(c) shows the edge disparity map using the BM approach. The edge disparity map is then clustered by the K-Means algorithm. Edges with large disparities will be classified as the background edges, while the small magnitudes as the reflection edges. Here we assume the background is closer to the camera then the reflection. It is just a change of notation if it is the other way round. Considering that some reflection and background edges can have similar depths and are unable to be accurately separated, we classify the edge disparity map into 3 classes instead of 2 as in (3).

$$\hat{E}_B^1 = (d^* > T_1) \times I_{ref}, \quad \hat{E}_R^1 = (d^* < T_2) \times I_{ref}, \quad (3)$$

where $\hat{E}_B^1$ and $\hat{E}_R^1$ are respectively the initial guess of the background and reflection edges, I_{ref} is the reference image (left image) with reflection. T_1 is the threshold for the background, and T_2 is the threshold for the reflection. These two thresholds are determined by the K-Means algorithm. As shown in (3), the edges with d^* in between T_1 and T_2 will be discarded, since their classification is error-prone. An example is shown in Fig. 2(d) and (e).

The missing background and reflection edges will be regenerated by a WGAN since GAN is known to be able to generate data following a given distribution and WGAN is more stable in training. Different from the traditional applications of GAN that a large degree of freedom is allowed for the image generated, we need the WGAN to give an output as close to the background edges as possible. The problem is more like an inverse problem than image generation. Thus, we suggest a structure similar to the conditional GAN [12] to ensure the output is close to the ground truth background edges. The operation of the network can be described by (4).

$$\hat{E}_B^2 = G_B\{\hat{E}_B^1, \hat{E}_R^1, E\}, \quad (4)$$

where E is the edges of the reference image. In (4), G_B is the generator of the WGAN. As shown in (4), the initial guess of the background and reflection edges $\hat{E}_B^1, \hat{E}_R^1$, and the edges of the reference image E, are fed to G_B to guide the generation of the complete background edges $\hat{E}_B^2$. An example of the generated background edges is shown in Fig. 2(g).

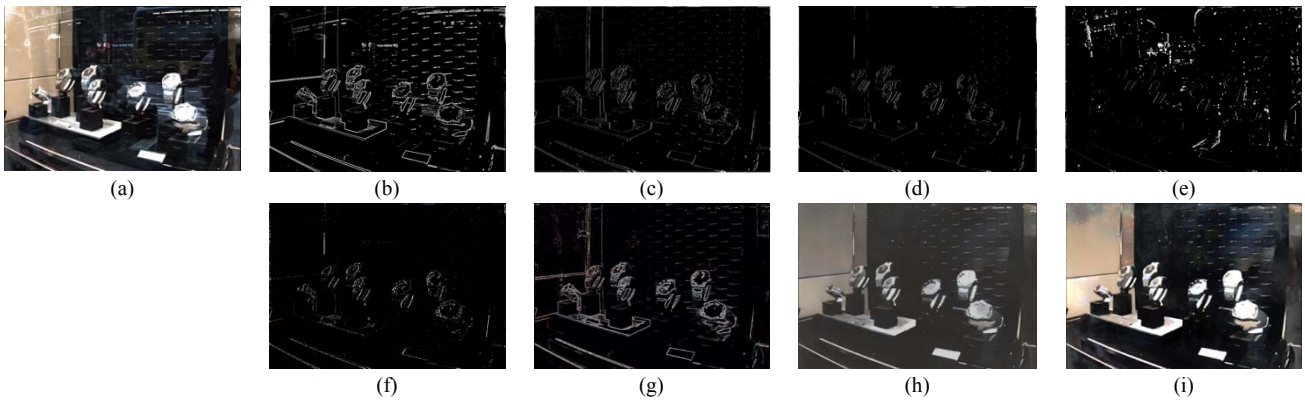


Figure 2. Results in each step. (a) The original image with reflection. (b) The strong gradients of the image. (c) Edge disparity map. (d) Background edge separated from the edge disparity map. (e) Reflection edge separated from the edge disparity map. (f) The initial guess of the background edges input to the network, i.e. $\hat{E}_B^1$. (g) The output of the WGAN, i.e. $\hat{E}_B^2$. (h) The reconstructed background B. (i) The final histogram aligned image.

The structure of the generator is shown in Fig. 3. It is similar to a U-net network with skip connections. They have been shown to be effective in the inverse problem in [13]. Different from the traditional GANs, two discriminators having the same structures as shown in Fig. 3 are used for the training of the WGAN. We train the generator G_B by updating its parameters using the gradient descent method to minimize the following function:

$$\sum_{i=1}^{m} \left\| G_B(z(i)) - E_B(i) \right\|_2^2 - \lambda_1 \left(D_B\left(G_B(z(i)) \right) + D_R\left(E - G_B(z(i)) \right) \right) \quad (5)$$

where m is the total number of training samples, E_B is the ground truth background edges, z is the input to the generator defined in (4). D_B and D_R are the discriminators for the background and reflection respectively, and λ_1 is the Lagrange multiplier. Note that the discriminator D_B is trained to give a large value if $G_B(z)$ gives an output close to the ground truth background edges and vice versa. The discriminator D_R is also trained to give a large value if $E - G_B(z)$ gives an output close to the ground truth reflection edges and vice versa. So, they combine to form a prior of $G_B(z)$ to regularize the first term in (5), which is a typical approach used in the inverse problem. Note that if $G_B(z)$ gives the background edges, $E - G_B(z)$ will give the reflection edges. Thus, the term $D_R\left(E - G_B(z)\right)$ strengthens the prior to regularize the optimization process.

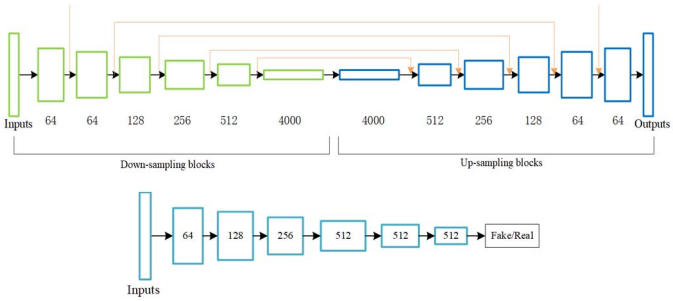


Figure 3. Structures of the generator and discriminators

The trained generator G_B is then used to train the discriminators D_B and D_R. Assume another m samples of the background and reflection image ground truths and m samples of the training images are obtained, the parameters of D_B and D_R are updated using the gradient ascent method (plus weight clipping) to maximize the following cost functions:

$$\sum_{i=1}^{m} D_B(E_B(i)) - D_B\left(G_B(z(i))\right), \quad (6)$$

$$\sum_{i=1}^{m} D_R(E_R(i)) - D_R\left(E - G_B(z(i))\right). \quad (7)$$

In (7), E_R is the ground truth reflection edges, which can be derived from the ground truth reflection image. The trained D_B and D_R are then used to train G_B again to obtain a better generator. The process repeats until converged.

For training the network, we synthesize the required training images with reflection by randomly adding two sets of light field (LF) images together with different weights. More specifically, we capture 318 sets of LF images and resize them to 256×256 pixels. Using the abovementioned approach, 112,225 images with reflection are synthesized as the training samples. We use the RMSprop solver [14] to train the generator and discriminators of the network with learning rates 2×10^{-4} and 2×10^{-5} respectively. The parameters λ_1 is set as 2.5×10^{-3}. The training and testing are both performed on a computer with Core i7 7820X CPU using a GTX 1080 Ti.

The estimated background edges $\hat{E}_B^2$ are then fed to the image reconstruction network given by [6] to reconstruct the background image. Similar to other edge-based image reconstruction methods, the reconstructed image is accurate only up to a scaling factor. We thus adopt the histogram specification method to align the distribution of the resulting background image to follow the original image. An example of the enhanced image after the histogram specification is shown in Fig. 2(i). It can be seen that the reflection is significantly removed while the color distribution of the image is very close to the original one.

III. Performance Evaluation

We compare the performance of the proposed approach with three difference learning-based reflection removal methods. The first comparing approach is the 5-view based method proposed in [6], while the other two are the single-image based methods [7-8]. For the qualitative comparisons, the testing images are obtained from real scenes using an LF camera. Two perceptual comparison results are shown in Fig. 4. Fig. 4(b) are the results using the proposed approach. It can be seen that the reflections in both images are effectively removed. As shown in Fig. 4, the proposed approach achieves a comparable performance as the 5-view-based approach. As to the results of the single-image based approaches shown in Fig. 4(d) and (e), it can be seen that only weak reflections can be partially removed. For the regions such as those cropped by the red boxes, single-image based approaches show their obvious limitations.

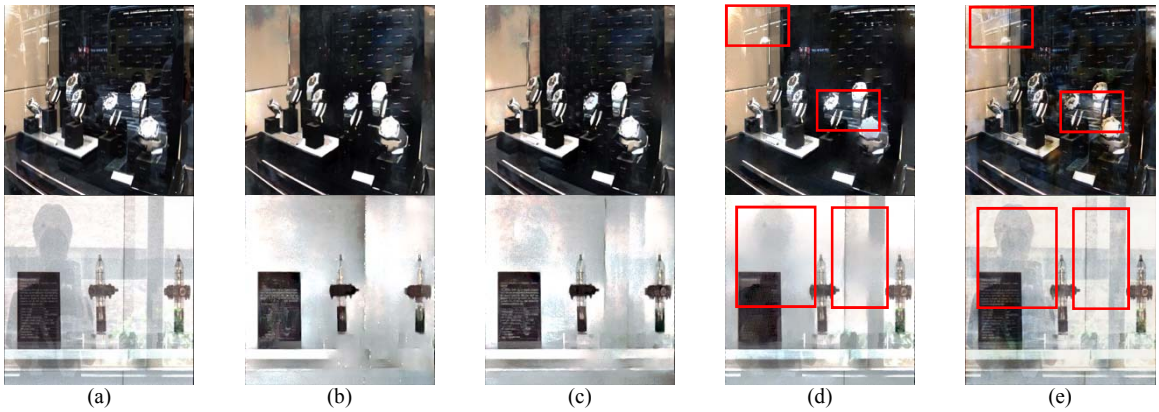
We have also quantitatively evaluated the reflection removal performance of different approaches using the synthesized images mentioned above. These testing images have not been used for the training of the network. Some of them and the corresponding ground-truth background are shown in Fig. 5. We use the ground truth background images as the references to calculate the average peak-signal-to-noise-ratio (PSNR), mean square error (MSE), and structural similarity (SSIM) of each approach. To avoid inconsistent intensity generated by each approach, we normalize the whole image by the Min-Max feature scaling method [15] to bring all pixel values into the range [0, 1]. The quantitative comparison results are shown in Table 1. It can be seen that the proposed approach outperforms the other two single-image based methods, and achieves a competitive result compared with the 5-view based approach while requiring a significantly simpler imaging hardware.

Table 1. Quantitative comparison results

Approach	**Proposed approach**	5-view based approach [6]	Single-image [7]	Single-image [8]
PSNR	**16.02**	16.62	13.99	11.01
MSE	**0.05**	0.04	0.05	0.06
SSIM	**0.61**	0.60	0. 52	0.34

IV. CONCLUSION

In this paper, the reflection removal problem using the stereoscopic images is considered. More specifically, we suggested a new background edge estimation algorithm that allows the edges of the background image to be effectively estimated. The new algorithm consists of a BM-based method for edge disparity estimation and a WGAN for regenerating the edges removed during the classification process. The comparison results show that the proposed method can outperform two recent single-image based reflection removal approaches; and achieve a competitive performance compared with a recent multiple-image based approach while having a much lower imaging hardware requirement. Further work of this study is being conducted on implementing the BM part with a learning-based approach to speed up the computation and improve the accuracy of the edge disparity estimation.



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

Figure 4. Qualitative comparison results. (a) Original images with reflections. (b) Results of the proposed approach. (c) Results of using a 5-view based approach [6]. (d) Results of using a single-image based approach [7]. (e) Results of using another single-image based approach [8].

Figure 5. Images used for quantitative evaluation. The second row shows the corresponding ground-truth background images.

REFERENCES

[1] Z. Chi et al. "Single image reflection removal ysing deep encoder-decoder network," *ArXiv* abs/1802.00094, 2018.

[2] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 29, no. 9, pp. 1647-29, 2007.

[3] 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, 2014.

[4] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," *in Proc, IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3193-3201.

[5] T. Li, D.P.K. Lun, Y.H. Chan, and Budianto, "Robust reflection removal based on light field imaging," *IEEE Trans. Image Process.,* vol. 28, no. 4, pp. 1798-1812, 2018.

[6] T. Li and D.P.K. Lun, "A novel reflection removal algorithm using the light field camera," *in Proc., 2018 IEEE Int. Sym. Cir. and Sys.*, 2018, pp. 1-5.

[7] T. Li and D.P.K. Lun, "Single-image reflection removal via a two-stage background recovery process," *IEEE Sig. Process. Lett.,* vol. 26, no. 8, pp. 1237-1241, 2019.

[8] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han and S. He, "Single image reflection removal beyond linearity," in *Proc. IEEE Int. Conf. Comput Vis. Pattern Recognit.,* 2019, pp. 3771-3779.

[9] 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 Int. Conf. Comput. Vis.*, 2017, pp. 3258-3267.

[10] R. Wan, B. Shi, H. Li, L. Duan, A. Tan and A. Kot, "CoRRN: Cooperative Reflection Removal Network," *IEEE Trans. Pattern Anal. Mach. Intell*. (early access), DOI: 10.1109/TPAMI.2019.2921574.

[11] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *in Proc. Int. Conf. Machine Learning*, 2017, pp. 214-223

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in Proc. *2017 IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967-5976.

[13] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit*, 2016, pp. 2536-2544.

[14] T. Tieleman and G. Hinton, "Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning,* vol. 4, pp. 26-31, 2012.

[15] R. Wan et al, "Benchmarking single-image reflection removal algorithms," *in Proc. IEEE Int. Conf. Comput. Vis.*, 2017.