# Single-Image Reflection Removal via a Two-Stage Background Recovery Process

Tingtian Li and Daniel P.K. Lun, *Senior Member, IEEE*

***Abstract*—The reflection problem often occurs when imaging through a semi-transparent material such as glass. It degrades the image quality and affects the subsequent analyses on the image. Traditional single-image based reflection removal methods assume the reflection is blurry. Deep neural networks (DNN) are then used to identify the blurry reflection and remove it. However, it is often that the blurry reflection still contains strong edges. They will be treated as the background and kept in the image. In this paper, we propose a novel two-stage DNN based reflection removal algorithm. In the first stage, we include a new feature reduction term in the loss function when training the network. Due to its strong reflection suppression ability, the reflection components in the image can be more effectively suppressed. However, it will also attenuate the gradient values of the background image. For recovering the background, in the second stage, we first estimate a reflection gradient confidence map based on the initial estimation result and use it to identify the strong background gradients. Then we use a generative adversarial network to reconstruct the background image from its gradients. Experimental results show that the proposed two-stage approach can give superior performance compared with the state-of-the-art DNN based methods.**

***Index Terms*—Image reflection removal, blind image separation, deep neural network**

## I. Introduction

In daily photography, it is often that the reflection of an unwanted scene is superimposed onto the background when imaging through a semi-transparent material such as glass. The reflection not only degrades the image visibility but also affects the subsequent analyses of the image. Mathematically, an image $I$ composed of the background layer $I_B$ and the reflection layer $I_R$ can be modeled as,

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

Removing the reflection layer from $I$ is a severally ill-posed blind separation problem because there are two variables needed to be solved from only one equation. Prior information about the background and reflection layers is needed to guide the separation processes to converge to the correct solutions. For instance, a manual labeling process is needed in [1] to indicate the positions of the background and reflection gradients in the image, but then the separation process will not be automatic. In [2-4], multiple images are used to utilize the motion or depth differences of the background and reflection layers. However, it requires the background scene to be absolutely static or motion blur will be introduced. We suggested in [5] using the light field images to solve the above problem, but special hardware is then required to capture the images. In [6], the background and reflection layers are assumed to follow different distributions. However, the method requires massive iterative optimizations on large scale matrices which take long computation time.

More recently, deep neural networks (DNN) which have gained dramatic successes in various areas are also applied to reflection removal [7, 8]. Inspired by [6], [7, 8] assume the background layer is in the depth-of-field while the reflection layer is not. Hence, the reflection is assumed to be blurrier than the background and has a distinct distribution different from the background. Based on this, the network can identify and remove the blurry reflection components. For example, [7] firstly uses a convolutional neural network (CNN) to distinguish the sharp background edges and then uses another CNN to reconstruct the image. [8] also trains a CNN to distinguish the blurry reflection components by minimizing the

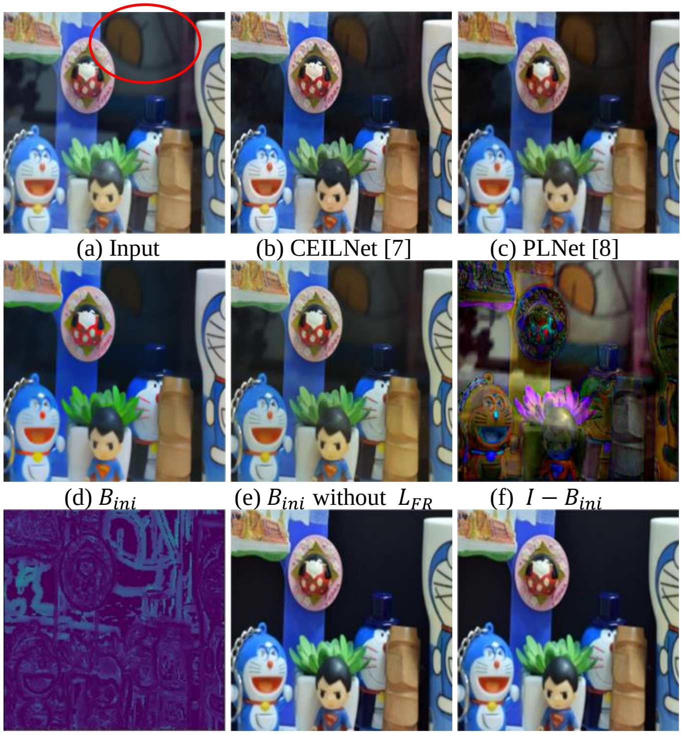

(a) Input (b) CEILNet [7] (c) PLNet [8]
(d) $B_{ini}$ (e) $B_{ini}$ without $L_{FR}$ (f) $I - B_{ini}$
(g) Confidence map $C_{rf}$ (h) Proposed $B_{final}$ (i) Ground truth

Fig. 1. Proposed reflection removal process. The reflection is blurry as it is out of focus, but the circled part contains strong gradient components that introduce much problem to traditional approaches.

This work was fully supported by the Hong Kong Polytechnic University under research grant RU9P. (Portions of) the research in this paper used the `SIR2' Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore. *(Corresponding author: Daniel P.K. Lun.)*

Tingtian Li, Daniel P.K. Lun are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. (e-mail: tingtianpolyu.li@connect.polyu.hk; enpklun@polyu.edu.hk).

VGG perceptual feature distance [9] compared with the ground truth. Although the assumption that the reflection layer is blurry is valid in many situations, we notice that some of the blurry reflection components can still have high gradient values. One example is shown in Fig 1(a) (the circled region). In this case, [7, 8] will mistakenly treat these strong gradients reflection edges as the background edges and leave in the resulting image as shown in Fig 1(b) and (c).

In this paper, we present a novel two-stage approach to remove the reflection using deep neural networks. In the first stage, we propose incorporating a feature reduction term in the loss function when training the CNN to enhance its reflection suppression ability. However, this term can also attenuate the background gradients and result in the blurry background. Therefore, we propose the second background refinement stage. We use the initial background estimation result to generate a confidence map for identifying the strong reflection and background gradients. Then, a generative adversarial network (GAN) is used to reconstruct the background image from the classified gradients. The GAN can also help in regenerating the background gradients that are accidentally removed in the first stage. Experimental results show that the proposed algorithm can give superior performance compared to other single-image deep learning based reflection removal methods.

## II. Initial Background Estimation with Feature Reduction Term

Perceptual features are widely used in deep-learning approaches for solving inverse problems [8-10]. Compared to the pixel-wise intensity, minimizing the perceptual feature distances can generate an image closer to the human perceptual expectation. The perceptual features can be obtained by extracting the intermediate layer features of a pre-trained network such as VGG-16, VGG-19 [11] trained on a large dataset. In fact, [8] also tries to remove the blurry reflection by minimizing the perceptual feature distance. Because this method highly depends on the assumption that the reflection components are blurry, it will fail when some parts of the reflection still show high gradient values. Just the perceptual feature distance is not enough to totally remove the reflection.

When an image $I_2$ is superimposed onto another image $I_1$, the resulting image $I$ will contain the textures from both $I_1$ and $I_2$. The superimposed image $I$ will contain more perceptual features than the original image $I_1$ or $I_2$. We argue that a good reflection removal process should also minimize the perceptual features in the resulting image. For the first stage of the proposed reflection removal algorithm, a CNN is trained with a loss function $L_{ini}$ as shown as follows:

$$L_{ini} = L_{rec} + L_{FR}; \tag{2}$$

$$L_{rec} = \sum\nolimits_{i=1}^{5} \lambda_1 \left\| \Phi_i\big(F_1(I)\big) - \Phi_i(I_B) \right\|_2^2 + \lambda_2 \| F_1(I) - I_B \|_1; \tag{3}$$

$$L_{FR} = \sum\nolimits_{i=1}^{3} \lambda_3 \left\| \Phi_i\big(F_1(I)\big) \right\|_1, \tag{4}$$

where $\Phi_i$ denotes the features at 'conv($i$_2)' layer of the pre-trained VGG-19 network on the ImageNet dataset [12]. $I_B$ is the ground truth background image. $\lambda_1$, $\lambda_2$, and $\lambda_3$ are the hyper-parameters, which are chosen as 3, 0.4 and 3, respectively, in our experiments. $F_1$ represents the proposed CNN, so $B_{ini} = F_1(I)$ gives an initial estimation of the background image. $L_{ini}$ in (2) consists of 2 loss functions $L_{rec}$ and $L_{FR}$. $L_{rec}$ serves to preserve the background. It is a weighted sum of the feature distance and pixel-wise distance from the background ground truth as shown in (3). Since the background images we used to train the network are all sharp and clear, $L_{rec}$ in effect guides the network to preserve the pixels or perceptual features come from the clear parts of the image. But if there exist some edges with strong gradients in the blurred regions that look like the background, the network will be confused and also preserve them. To solve the problem, we propose to incorporate a feature reduction term $L_{FR}$ as shown in (4) when training the CNN. It gives the total feature magnitudes of the first few layers of a VGG-19 network with $B_{ini}$ as the input. It serves to minimize the low level perceptual features of $B_{ini}$. Since $L_{FR}$ will lead to the suppression of all low level features and $L_{rec}$ will try to preserve only the background features, it ends up that the reflection features will be suppressed more comparing to the background features. More importantly, for edges with strong gradients in the blurred regions, $L_{FR}$ will tend to guide the network to remove them. They will not be preserved by $L_{rec}$ since these edges are often isolated (their neighbors are blurred) and thus do not have much semantic meaning. Fig. 1(d) and (e) show a comparison between $L_{ini}$ with and without $L_{FR}$. We can see that if $L_{FR}$ is not included, the result in Fig. 1(e) is similar to Fig. 1(b) and (c). There are obvious reflection edges remaining in the result. On the contrary, the result of using $L_{FR}$ in Fig. 1(d) shows much weaker reflection edge residues. But as expected, $L_{FR}$ will also lead to the reduction of low-level background features, although is not as much as to the reflection. It is noticed that $B_{ini}$ is blurrier than the ground truth and may also be blurrier than Fig. 1(b) and (c). In the next section, we will discuss the second stage of the proposed algorithm for refining the background image.

## III. Background Refinement at The Second Stage

The reduction of low-level features in $B_{ini}$ renders the attenuation of its gradient values. Interestingly, it provides us useful information to identify the strong gradients of the background and reflection layers. In fact, as discussed in [3, 5], the background layer can be reconstructed from its strong gradients, while those flat regions with weak gradients can be easily inferred by the networks or optimization processes. Now, let us consider the residue of the initial background estimate, i.e. $(I - B_{ini})$. It contains mainly the reflection layers plus the attenuated background gradients as shown in Fig. 1(f). Comparing $(I - B_{ini})$ with $B_{ini}$, the attenuated background gradients in $(I - B_{ini})$ overlap with the background gradients in $B_{ini}$. And according to the gradient independence property [3-5], the strong gradients of the background and reflection layers seldom overlap since they are often uncorrelated. It means that at the positions where the strong reflection gradients in $(I - B_{ini})$ are found, we will not find any strong background gradients in $B_{ini}$. Based on the above, we define a confidence map for identifying the strong reflection gradients as follows:

$$C_{rf} = log\left(\frac{G_{I-B_{ini}}}{G_{B_{ini}} + \varepsilon} + 1\right) \cdot M, \tag{5}$$

where $G$ represents the gradient magnitude, $\varepsilon$ is a very small constant. $M$ is a mask which has the value of 1 for those pixels in $I$ with the Sobel gradient magnitude larger than 1, and 0 otherwise. It masks out only the positions in $I$ where strong gradients are found for the subsequent operations. As mentioned above, at the positions where $G_{I-B_{ini}}$ contains strong reflection gradients, $G_{B_{ini}}$ will have small or even 0 values. And at the positions where $G_{I-B_{ini}}$ contains the attenuated background gradients, $G_{B_{ini}}$ will have the original background gradients which have larger values. Thus, only the reflection strong gradients will have high confidence values in $C_{rf}$ as shown in Fig. 1(g). Then we run a K-means clustering process ($K = 2$) on this confidence map. It will generate an adaptive threshold $\xi$ to classify the values in $C_{rf}$ into two groups. The reflection strong gradients $E_R$ and background strong gradients $E_B$ can then be identified as follows:

$$E_R = E_I \cdot (C_{rf} > \xi); \;\; E_B = E_I \cdot (C_{rf} < \xi), \tag{6}$$

where $E_I$ denotes those pixels in $I$ whose gradient magnitudes above 1. We concatenate the image $I$ with $E_B$ and $E_R$ to form the input z and sent to a new network $F_2$ for background reconstruction. A loss function is defined as follows:

$$L_2 = \sum_{i=1}^{5} \lambda_1 \left\|\phi_i(F_2(z)) - \phi_i(I_B)\right\|_2^2 + \lambda_2\|F_2(z) - I_B\|_1 - \lambda_4 D(F_2(z)). \tag{7}$$

Similar to $L_{rec}$ in the first stage, the first two terms are used to reconstruct the background. Because $E_B$ and $E_R$ may contain some outliers, we also use an adversarial term $-\lambda_4 D(F_2(z))$ to guide the results to follow the distribution of natural images. (7) can be implemented using a GAN, where $D$ is the discriminator for measuring the similarity between the inferred background $F_2(I)$ and the ground truth background $I_B$ . $\lambda_4$ is a hyper-parameter which is chosen as 0.05 in our experiments. The discriminator $D$ will show high values when $F_2(I)$ follows the distribution of natural images. This discriminator can be jointly trained by minimizing the following loss function,

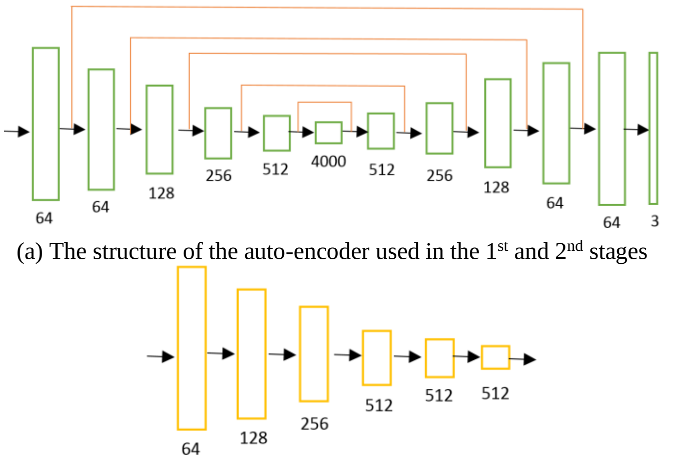
$$L_{adv} = D(F_2(z)) - D(I_B). \tag{8}$$

Fig. 1(h) shows an example of the final result $B_{final}$. It can be seen that the final background becomes shaper without reflection residues.

## IV. Experiments and Results

### A. Network Structure

The networks we used for the proposed algorithm are conventional. For the first stage, a U-net like auto-encoder as shown in Fig 2(a) is used to generate the initial background estimation. U-net has been widely used in solving inverse problems [13-15]. In our experiment, the encoder contains 6 levels of stride-two convolutional layers, followed by batch normalization layer and ReLU. The decoder part consists of also 6 levels of deconvolutional layer, followed by batch normalization layer and leaky ReLU. We also concatenate the features at the encoder side to the decoder side at each level for increasing the resolution of the results [15]. For the second stage, the structure of the generator network is the same as the auto-encoder used in stage 1. The discriminator network is relatively simple as shown in Fig. 2(b). It is composed of six blocks of stride-two convolutional layers, batch normalization layers and leaky ReLU. The output of the discriminator is a scalar value indicating the judgement that the perceptual features are real or fake.



(a) The structure of the auto-encoder used in the 1st and 2nd stages

(b) The structure of the discriminator used in the 2nd stage

Fig. 2. The structures of the networks used in the proposed algorithm.

### B. Training Data Preparation

For training the networks, we synthesize images with reflection using the images in the VOC2012 dataset [16]. The synthetization strategy is similar to [7, 8]. A training sample is synthesized by superimposing one image serving as reflection on another image serving as the background. Images in the dataset are mixed together randomly so that many training samples can be obtained. We simulate the blurring effect of the reflection layer by smoothing the reflection images before adding them to the background images. We also simulate the possible ghost effect [17] by convolving the reflection images with a kernel with two very close impulses. The synthesized images are then resized to $256 \times 256$. Rotation and flipping are also used for data augmentation. The networks of the first and second stages are trained sequentially for avoiding the function loss of the first stage due to overfitting. We use the RMSprop solver [18] to train $F_1$, $F_2$ and $D$. The learning rates of them are $2 \times 10^{-4}$, $2 \times 10^{-4}$ and $2 \times 10^{-5}$ respectively. The training

Table I.
The Average PSNR of Different Methods

| Method | PSNR |
|---|---|
| CEILNet [7] | 21.75 |
| PLNet [8] | 20.28 |
| Proposed w/o the adversarial term in (6) | 22.25 |
| Proposed | 23.01 |

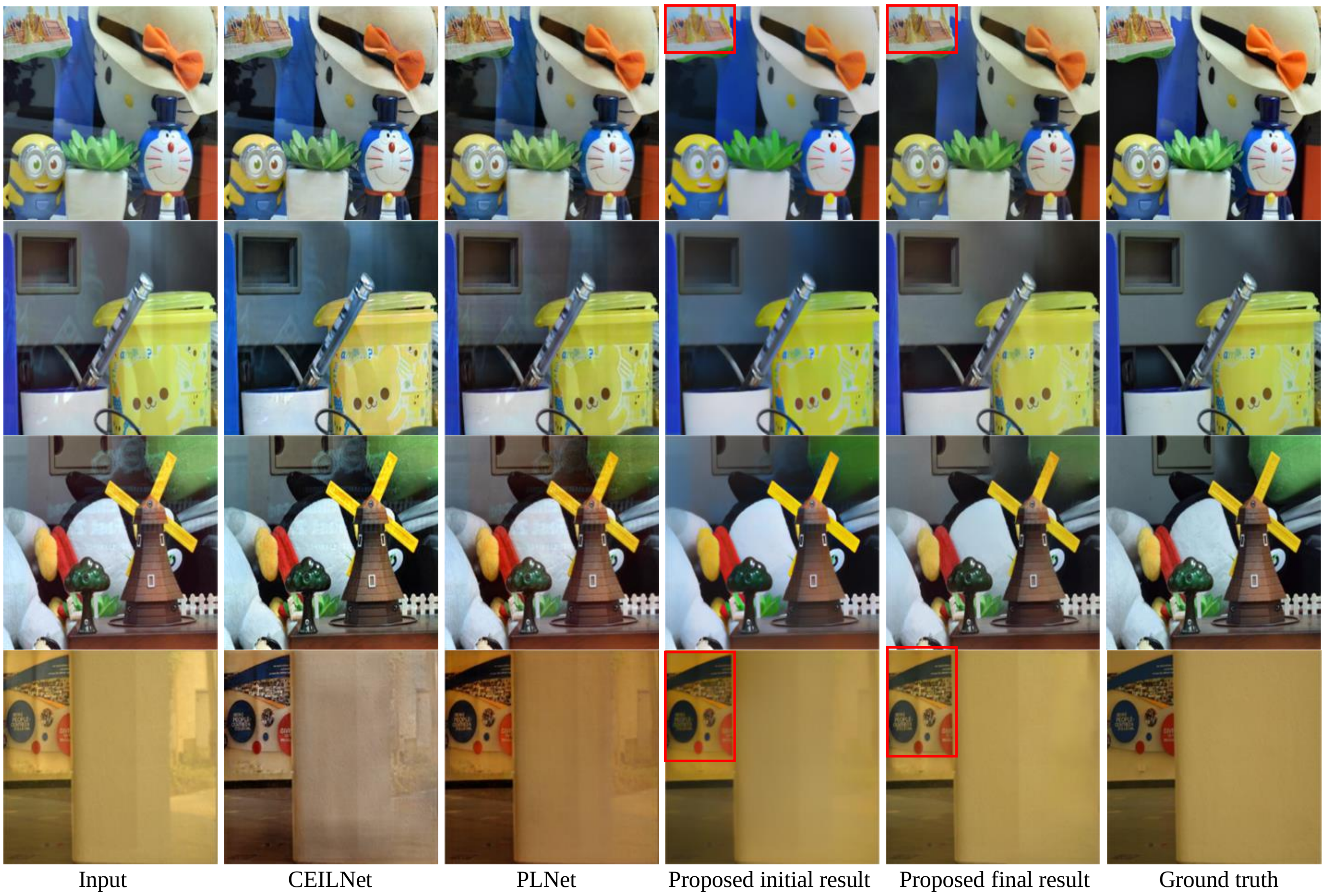


Fig. 3. Reflection removal using different approaches on the images from a benchmark dataset SIR2 [19].

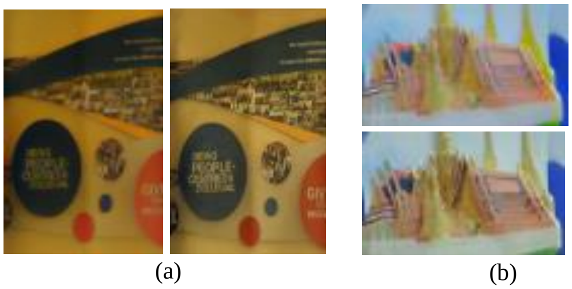


Fig. 4. Blow-ups of the red boxes in Fig. 3 to compare the initial and final results. (a) Left: initial, right: final; (b) top: initial, bottom: final.

and testing are performed on a computer using a GTX 1080 Ti.VIA

## C. Evaluation and Comparisons

For evaluating the performance of the proposed approach, we conduct a series of quantitative and qualitative comparisons with two recent DNN-based single-image reflection removal methods: CEILNet [7] and PLNet [8]. They are implemented with the source codes published on their websites. The parameters used are the same as the ones in their codes. We test all methods using the images from a benchmark dataset SIR2 [19]. It consists of real images with reflection and their ground truth background. The average PSNR of the results with respect to the ground truth images are shown in Table I. We can see that the proposed method significantly outperforms the other two methods. We also show the performance of the proposed method with and without the adversarial term in (7). It shows that the adversarial term can further improve the performance by guiding the resulting image to follow the distribution of natural images. It can regenerate the missing background gradients due to the estimation errors in $E_B$ and $E_R$.

The qualitative comparison results are shown in Fig. 3. It can be seen that both CEILNet [7] and PLNet [8] have obvious reflection residues in their results. They mainly come from the strong edges of the reflection layers which have high gradient values. On the contrary, the initial stage of the proposed algorithm can better suppress the reflection components as shown in Fig. 3, although it can also blur the background as shown in Fig. 4. We make use of the information provided in the initial result to refine the background estimates in the second stage. It has the best quality as can be seen in Fig. 3 and is much clearer than the initial estimates as shown in Fig. 4.

# V. Conclusion

In this paper, a novel two-stage reflection removal algorithm using deep neural networks (DNN) is proposed. The new method can fully remove the reflection residues which often appear in the traditional methods when the reflection also contains strong gradient components. Our experimental results have demonstrated the superior performance compared to other state-of-the-art DNN-based methods. The proposed algorithm is particularly suitable to images with blurred reflection, which are often encountered in daily photography.

## References

[1] 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, 2007.

[2] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, p. 79, 2015.

[3] Y. Li and M. S. Brown, "Exploiting reflection change for automatic reflection removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2432-2439.

[4] X. Guo, X. Cao, and Y. Ma, "Robust separation of reflection from multiple images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2187-2194.

[5] T. Li, D. P. K. Lun, Y. H. Chan, and Budianto, "Robust Reflection Removal Based on Light Field Imaging," *IEEE Trans. Image Process.*, 2018.

[6] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2752-2759.

[7] 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, p. 4.

[8] X. Zhang, R. Ng, and Q. Chen, "Single Image Reflection Separation with Perceptual Losses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[9] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694-711.

[10] D. Engin, A. Genc, and H. Kemal Ekenel, "Cycle-Dehaze: Enhanced CycleGAN for Single Image Dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 825-833.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.

[13] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, pp. 4509-4522, 2017.

[14] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536-2544.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241.

[16] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 (voc2012) development kit," *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 2011.

[17] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3193-3201.

[18] 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," *COURSERA: Neural networks for machine learning*, vol. 4, pp. 26-31, 2012.

[19] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.