ABSTRACT

Removing undesired reflections from a photo taken in front of a glass is of great importance for enhancing visual computing systems’ efficiency. Previous learning-based approaches have produced visually plausible results for some reflections type, however, failed to generalize against other reflection types. There is a dearth of literature for efficient methods concerning single image reflection removal, which can generalize well in large-scale reflection type. In this study, we proposed an iterative gradient encoding network for single image reflection removal. Next, to further supervise the network in learning the correlation between the transmission layer features, we proposed a feature co-occurrence loss. Extensive experiments on the public benchmark dataset of SIR2 demonstrated that our method can remove reflection favorably against the existing state of the art method on all imaging settings, including diverse backgrounds. Moreover, as the reflection strength increases, our method can still remove reflection even where other state of the art methods failed.

Index Terms— SIRR, feature co-occurrence loss, iterative gradient encoding

1. INTRODUCTION

How to obtain a reflection-free image taken through the glass is of great interest to computer vision researchers. Removing the undesired reflection enhances the target object’s visibility and benefits various computer vision tasks, such as image classification, segmentation, and object detection. The initial efforts of reflection removal employed multiple images to disentangle the reflection from the transmission [1], [2], [3]. More recently, the endeavor to solve the more common and practically significant scenario of a single input image has received a lot of appreciation [4], [5], [6]. However, single-image reflection removal is a challenging process because of the ill-posed nature of the problem [7]. To deal with the problem’s ill-posedness, the recent learning-based method has utilized different auxiliary information as prior and constraint [8], [9], [10], [11]. Among them, one school of researchers had tried to exploit auxiliary information embedded in the image’s gradient [12], [13]. For example, in [9] Wan et al. proposed to use an auxiliary network to restore the gradient of the corrupted image and fused the features of the gradient network with the image restoration network. In a different work, M Ikehara et al.[12] proposed a feature space gradient constraint loss. In [13] Ket et al. proposed an MMD loss based on the higher-order statics of image gradient. Together with other learning-based methods, these methods perform adequately when the effect of reflection is low; however, they fail miserably if the intensity of reflection increases. Moreover, due to different complexities such as insufficient training data, varying imaging conditions, diverging scene content, and little physical understanding of this problem, these methods often fail to remove some reflection but perform sufficiently for another reflection type. In this study, we proposed a novel iterative gradient encoding network for single image reflection removal. Our proposed iterative method first computes the gradient of the estimated transmission layer of the current iteration and utilizes it to estimate the transmission layer in the next iteration. We have employed a multi-scale feature fusion scheme to use the feature of gradient image. Note, our proposed method is different from other iterative methods, as we have not used the estimated transmission as an input to the network in the next iteration; only a subpart of the network is iterated, holding the feature extracted from the mixture image fixed.

Next, we reintroduced the well-known style loss [14] as feature co-occurrence loss for reflection removal. The correlation between the features of the transmission layer and the reflection layer provides useful information for separating these layers. Style loss is well known for capturing correlation between different features, yet this consequential but straightforward loss is unused in reflection removal. To the best of our knowledge, this is the first study to adopt style loss as feature co-occurrence loss for reflection removal. We evaluated our proposed method in the recent benchmark data set of SIR2 [15]. This data set contains real-world reflection images with diverse imaging settings and backgrounds. Extensive experiment on this data set has shown our method can remove reflection more efficiently than the current state of the art in all imaging conditions. Moreover, our iterative methods have helped to remove strong reflections from the background where most of the state of the art method failed.
2. PROPOSED METHOD

Gradient provides vitally important clues for separation of reflection from the background. In general, the gradient of the transmission layer contains larger values, and the reflection layer is with smaller gradient values. So the pixel belonging to the transmission layer can be more easily differentiated in the gradient domain. In this work, we proposed an iterative gradient encoding network to encode the estimated transmission layer’s gradient image and use these features to identify the pixel location belonging to the transmission layer in the next iteration. Our proposed iterative gradient encoding network is shown in Figure 1. Our proposed network comprise of three main components: image encoding sub-network $\phi$, gradient encoding sub-network $\theta$, and iterative image reconstruction network $\psi$. The iterative reconstruction network’s output is the estimated transmission layer $\hat{T}$. At an iteration step $t$, the output $\hat{T}(t)$ is given by,

$$\hat{T}(t) = \psi\left(\phi(I); \theta\left(\nabla \cdot \hat{T}(t-1)\right)\right)$$

Where $I$ is the original mixture image and $\nabla$ is the gradient operator. We initialize the $\hat{T}(0)$ as $I$. Note, the image encoding network’s output and input are fixed across all iterations; at every iteration, these encoded features are being mixed with the updated gradient features. We speculated that the gradient would be more prominent at every iteration. The gradient features evolution will guide the iterative reconstruction network to identify the pixel belonging to the transmission and interpolate the remaining pixel based on these pixels. To effectively learn the evolution of gradient features as well as the long-range dependency among gradient features and image features, we have utilized the well-known convolutional LSTM unit.

The efficacy of multi-scale representation for reflection removal has already been well studied in the previous literature [9], [16]. Inspired by these works, we have used a novel multi-scale feature fusion scheme to fuse gradient image features with the original image. As shown in Figure 1, we have used three different scales for feature fusion; at every scale, first, the features of the image encoder are mixed with the interpolated feature by a convolution layer, these features are concatenated with the updated gradient feature and then forwarded to a two-layer convolutional cell for interpolation followed by upsampling. In every convolutional layer and convolutional LSTM cell, we have used $3 \times 3$ kernel.

**Feature Co-Occurrence Loss:** In the mixture image, objects from transmission layers and reflection layers are overlaid with each other. However, in the original transmission layer, this type of superposition does not happen. In this study we proposed to use feature co-occurrence loss to regulate this type of superimposition. It is defined as:

$$\mathcal{L}_f = \sum_{l} \left| \left| G(\zeta^l(\hat{T})) - G(\zeta^l(T)) \right| \right|_2^2$$

with Gram matrix $G(F) = FF^T \in \mathbb{R}^{n \times n}$, and $\zeta^l(I) \in \mathbb{R}^{n \times m}$ is the feature activation at the $l^{th}$ pre-trained VGG layer [17] with $n$ features of length $m$. Note, our objective is not to transfer the original transmission image’s style content but to restrict unrealistic feature co-occurrence in the estimated transmission image by forcing the correlation among the feature extracted from the estimated transmission image to resemble the feature correlation extracted from the original transmission image.

**Other Loss Function:** We have also used Mean Absolute Error/ L1 Loss ($\mathcal{L}_1$), perceptual loss ($\mathcal{L}_p$), and adversarial loss ($\mathcal{L}_{adv}$) between real transmission layer ($\hat{T}$) and estimated transmission layer ($\tilde{T}$) to train our network. So, our total loss function $\mathcal{L}_t$ is given as:

$$\mathcal{L}_t = \mathcal{L}_1(\hat{T}, T) + \lambda_1 \mathcal{L}_p(\hat{T}, T) + \lambda_2 \mathcal{L}_f(\hat{T}, T) + \lambda_3 \mathcal{L}_{adv}(\hat{T}, T)$$

We empirically set the values of weight $\lambda_1, \lambda_2, \lambda_3$ as 0.1, 0.1, 0.5. Details about these loss functions and architecture details of discriminator network is given in the supplementary material.
3. EXPERIMENTAL DETAILS:

We have used the recently proposed Reflection Image Dataset (RID) [13] as the training set. Also followed the similar procedure as given in [13] to synthesize the reflection images. For training our network, we used the whole images instead of the patch-based training strategy. For optimization, we used Adam optimizer with a batch size of 16. The remaining training details are given in the supplementary material. For computing gradient, we have used the following three filter on each of the RGB channels,

\[ f_1 = [-1 1], f_2 = [-1 1]^T, f_3 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \]

For evaluation of our proposed method, we have used the recently proposed benchmark data set SIR^2 [15]. This data set contains real-world reflection images with diverse backgrounds and imaging conditions. For comparison we considered following state of the art methods, PNet [18], ERRNet [19], GCNet [12], IBCLN [6], CoRRN [13]. For a fair comparison, we have used the codes and trained model provided by the original authors.

4. RESULT AND DISCUSSION

In this section, we first show the efficacy of the proposed iterative reconstruction network. In Figure 2 we have given an example reflection image taken from wild screen subset of SIR^2 dataset, and estimated transmission image at different iteration is shown on the right-hand side. The effect of the reflection is becoming more debilitated as the number of iteration is increasing. In the rightmost image of Figure 2 we have given the result of reflection removal of the same network but trained without feature co-occurrence loss. The reflection in the shutter of the left side shop is still present. Noticeably, ERRNet, IBCLN completely failed to remove the reflection effect. Next, we considered reflection images from the postcard subset. This subset contains outdoor reflection images with complicated environment. Here we can see reflection removal result of our method is favorably better than the other state of the art methods. Notably, ERRNet, IBCLN, completely failed to remove the reflection effect.

For evaluation of our proposed method, we have used the recently proposed benchmark data set SIR^2 [15]. This data set contains real-world reflection images with diverse backgrounds and imaging conditions. For comparison we considered following state of the art methods, PNet [18], ERRNet [19], GCNet [12], IBCLN [6], CoRRN [13]. For a fair comparison, we have used the codes and trained model provided by the original authors.

We have added more results in the supplementary material with different imaging conditions and backgrounds to observe the benefit of the iterative reconstruction network and feature co-occurrence loss. Next, following the de facto practice in the literature, we have compared our reflection removal result with other state of the art methods objectively in Table 1 in terms of PSNR and SSIM. Our method has achieved the best PSNR and SSIM among the current state of the art methods like PNet, ERRNet, and GCNet, CoRRN. Next, we performed a blind reader study for subjective evaluation. We asked five volunteers to rate twenty randomly selected images from the SIR^2 dataset on a five-point scale regarding reflection removal and structure preservation. The mean opinion score (MOS) of the study is presented in Table 1; our method has achieved the best MOS among all state-of-the-art methods.

Next, we present some of the reflection removal results for subjective evaluation in Figure 3. In the top row of Figure 3 an example image from the wild screen subset is shown. The corresponding estimated transmission image of the different methods is shown on the right side. Wild screen set contains outdoor reflection images with complicated environment. Here we can see reflection removal result of our method is favorably better than the other state of the art method. Noticeably, ERRNet, IBCLN, completely failed to remove the reflection effect.

![image](image.png)

Fig. 2. Visualization of results at different iteration evaluated on the Wild Screen Dataset of SIR^2. The estimates of transmissions become increasingly more accurate with increase in the iteration. More results are on the supplementary material.

| Method   | Metric | PSNR | SSIM | Reflection Removal | Structure Preservation |
|----------|--------|------|------|--------------------|------------------------|
| PNet [18] |        | 20.33 | 0.833 | 3.9                | 3.5                    |
| ERRNet [19] |      | 23.66 | 0.870 | 3.6                | 3.8                    |
| GCNet [12] |       | 22.82 | 0.926 | 3.2                | 3.3                    |
| IBCLN [6]  |        | 24.32 | 0.884 | 3.3                | 3.5                    |
| CoRRN [13] |       | 24.19 | 0.903 | 3.8                | 3.4                    |
| Proposed   |        | 24.52 | 0.927 | 4.2                | 4.0                    |

Table 1. Quantitative evaluation results using PSNR, SSIM and Mean Opinion Score (MOS). The result of objective evaluations are obtained by averaging the metric scores of all images from SIR^2 dataset.

For evaluation of our proposed method, we have used the recently proposed benchmark data set SIR^2 [15]. This data set contains real-world reflection images with diverse backgrounds and imaging conditions. For comparison we considered following state of the art methods, PNet [18], ERRNet [19], GCNet [12], IBCLN [6], CoRRN [13]. For a fair comparison, we have used the codes and trained model provided by the original authors.

We have added more results in the supplementary material with different imaging conditions and backgrounds to observe the benefit of the iterative reconstruction network and feature co-occurrence loss. Next, following the de facto practice in the literature, we have compared our reflection removal result with other state of the art methods objectively in Table 1 in terms of PSNR and SSIM. Our method has achieved the best PSNR and SSIM among the current state of the art methods like PNet, ERRNet, and GCNet, CoRRN. Next, we performed a blind reader study for subjective evaluation. We asked five volunteers to rate twenty randomly selected images from the SIR^2 dataset on a five-point scale regarding reflection removal and structure preservation. The mean opinion score (MOS) of the study is presented in Table 1; our method has achieved the best MOS among all state-of-the-art methods.
examples with different settings in a controlled environment. In the first example, we took a reflection image taken by the camera with an aperture size of F11 and shutter speed 1/3s. In this example, we can see, CoRRN is the least performer in terms of reflection removal. Non arguably, our method has removed the reflection effect most efficiently. Next, we consider another image of the same object but taken by the camera with an aperture size of F32 and shutter speed 3s. Big aperture size makes the reflection more vital [15]; as shown in 3rd-row images of Figure 3. This time ERRNet, GCNet, IBCLN again failed to remove reflection; however, they decently removed reflection in the previous setting with low reflection strength. The reflection removal result of CoRRN and PNet is lower than the proposed method for this example also. Our iterative method makes it possible to remove strong reflection from the background by iteratively making the background more accurate than the previously estimated one. To further concrete our claim, we again took another challenging example from the solid object subset. Here the reflection image is taken through a thick glass of 5mm. Thick glass creates a ghosting cue effect in the reflection image [5]. We can see that most of the state of the art methods failed to remove the ghosting cue effect, whereas our method has acceptably removed reflection from the image. Next, we took another reflection image from the wild screen subset with a very strong reflection effect. Removing reflection from these mixture images is challenging due to the complex type of reflection; our method has still performed favorably better than the other methods. We found different methods perform satisfactorily in removing specific reflection types but exhibited limitations when the kind of reflection changes. These methods either considered a distinct image formal model while designing the loss function, regularization constraint, and choosing the prior or are completely data-driven. Whereas, our data-driven method utilized special features of the gradient images, but without any implicit consideration of the reflection formation model, which makes our method better generalized over diverse reflection images. In the supplementary material, we have given more results of reflection removal of our method.

5. CONCLUSION

This paper presents a novel method for single image reflection removal. Specifically, we proposed an iterative gradient encoding network and a feature co-occurrence loss. Unlike the conventional pipeline, we did not use any gradient inference network but used an encoding network to use the features of the gradient for understanding the pixel location of the transmission layer. Comprehensive experiments on the benchmark dataset of SIR$^2$ validates that our method is more competent in removing strong reflection than the existing state of the art method. We have evaluated our method in reflection images taken in a controlled environment with different settings. Our method satisfactorily performed in all settings, unlike the previous method, which failed to perform in all settings with the same competency.
6. REFERENCES

[1] Xiaojie Guo, Xiaochun Cao, and Yi Ma, “Robust separation of reflection from multiple images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2187–2194.

[2] Bernard Sarel and Michal Irani, “Separating transparent layers through layer information exchange,” in European Conference on Computer Vision. Springer, 2004, pp. 328–341.

[3] Richard Szeliski, Shai Avidan, and Padmanabhan Anandan, “Layer extraction from multiple images containing reflections and transparency,” in Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662). IEEE, 2000, vol. 1, pp. 246–253.

[4] Yu Li and Michael 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.

[5] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman, “Reflection removal using ghosting cues,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3193–3201.

[6] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft, “Single image reflection removal through cascaded refinement,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3565–3574.

[7] Anat Levin and Yair Weiss, “User assisted separation of reflections from a single image using a sparsity prior,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 9, pp. 1647–1654, 2007.

[8] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf, “A generic deep architecture for single image reflection removal and image smoothing,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3238–3247.

[9] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot, “Crrn: Multi-scale guided concurrent reflection removal network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4777–4785.

[10] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi, “Reflection separation using a pair of unpolarized and polarized images,” in Advances in neural information processing systems, 2019, pp. 14559–14569.