Learned Dual-View Reflection Removal
— Supplementary Material —

Simon Niklaus
Adobe Research

Xuaner (Cecilia) Zhang
UC Berkeley

Jonathan T. Barron
Google Research

Neal Wadhwa
Google Research

Rahul Garg
Google Research

Feng Liu
Portland State University

Tianfan Xue
Google Research

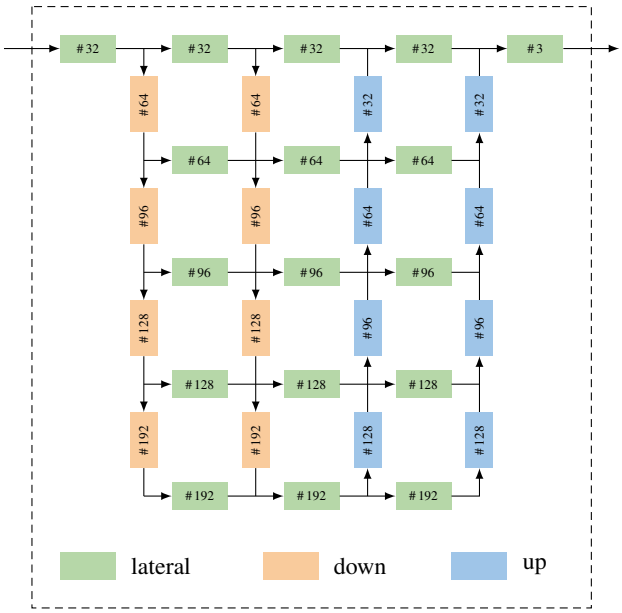

Figure 1: The architecture of our synthesis network which takes I_1 and $I_{2\to1}$ and predicts T_1. Please see Figure 2 for details of each block. The annotations are the number of output channels for the last convolution layer of each block.

1. Synthesis Network

We use a GridNet [1] with the modifications from Niklaus *et al.* [5] for our synthesis network. It consists of five rows and four columns where the first two columns perform downsampling and the last two columns perform upsampling. Please see Figure 1 for an illustration of the GridNet architecture that we employed as well as Figure 2 for details about the composition of each building block.

2. Exposure Adjustment

The image formation model for our dual-view dataset reduces the brightness of the transmissive layer by α where α is randomly chosen. As such, we supervise our image synthesis model to estimate $\alpha \cdot T$ instead of T which, as shown in Figure 3, has the side-effect of the synthesized images being darker than desired if there are many reflections. To account for this, we employ a simple auto-exposure approach that post-processes the prediction. Specifically, we apply an additional gain to make sure that the 60-percentile brightness of the image is no less than 0.5, and 95-percentile brightness is 1.0 (the image intensity range is $[0, 1]$). This is similar the auto exposure algorithm in [6]. This simple post-processing correctly adjusts the brightness of the synthesis result and yields less-dark and overall more-pleasant results.

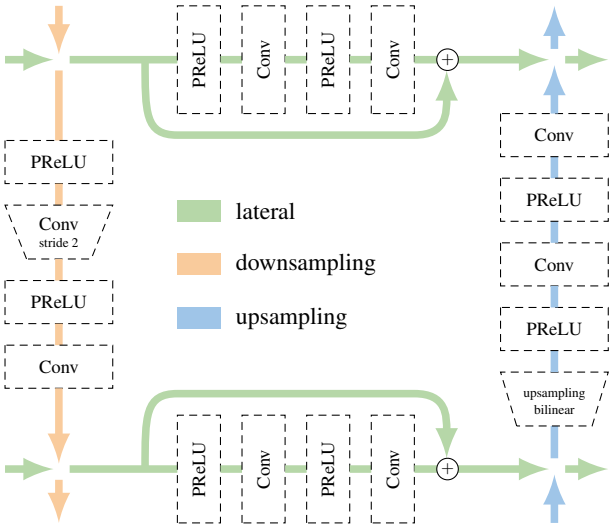

Figure 2: An illustration of the building blocks that we used in our synthesis network. We use parametric rectified linear units initialized with $\alpha = 0.2$ as activation functions.

3. Directional Invariance

Our camera setup captures five viewpoints at a time, with one camera in the center and one camera in each orthogonally adjacent direction (left, right, up, down). We use the center as reference and the surrounding viewpoints as the second input when employing dual-view reflection removal. As such, we are able to get four different results per scene capture.

Before Auto-exposure After Auto-exposure

Figure 3: Our synthesis model tends to predict relatively dark dereflection results due to the image formation model that we used when creating the training data. We ameliorate this by using a simple auto-exposure post-processing step.

		rendered test set			real-world test set		
	images used	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
center + left	2	*N/A*	*N/A*	*N/A*	22.86	0.766	0.103
center + right	2	*N/A*	*N/A*	*N/A*	22.86	0.766	0.103
center + up	2	*N/A*	*N/A*	*N/A*	22.82	0.765	0.105
center + down	2	*N/A*	*N/A*	*N/A*	22.76	0.765	0.105

Table 1: Evaluation of our approach with respect to the position of the second view. Please note that there are no metrics (N/A) for our rendered test set since it only comes with one (randomly positioned) second view per sample.

We reported averages in our main paper and show the per-direction metrics of our approach in Table 1. This evaluation demonstrates that our dual-view approach is able to generate high-quality results regardless of and independent from the directional location of the second input viewpoint.

4. Additional User Study

Unfortunately, almost no multi-view reflection removal papers provide reference implementations for their proposed approach. We were thus forced to limit the comparison in our main paper and only included the multi-view approach from Li and Brown [3]. For a more broad comparison, we conducted an additional user study in which we compare to more multi-image dereflection methods based on the results that the authors provided in their papers or websites. Specifically, we include results from [2, 3, 4, 7, 10] on footage from [3, 7, 10] and conducted an A/B user study with 19 participants. There are 17 sequences in total with between 5 and 9 frames each. We chose the first and middle frames as input to our approach whereas the baselines use all available frames. We provided the participants with a comparison tool where they could switch between our result and the result of a baseline method (the order of comparisons was randomized for each participant), and each participant was asked

		test images from [3]	test images from [7]	test images from [10]
	images used	prefer ours	prefer ours	prefer ours
Guo *et al.* [2]	5 – 9	98.0%	98.3%	99.3%
Li & Brown [3]	5 – 9	92.1%	82.5%	92.1%
Liu *et al.* [4]	5 – 9	78.3%	42.1%	40.8%
Sinha *et al.* [7]	5 – 9	*N/A*	36.8%	*N/A*
Xue *et al.* [10]	5 – 9	*N/A*	*N/A*	28.3%

Table 2: Results from an additional A/B user study where participants compared the results from out method with the ones from several multi-image derefelction baselines.

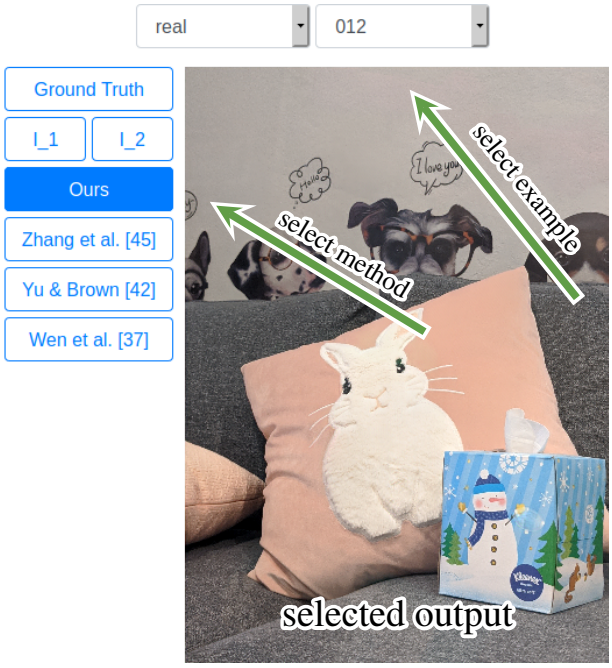

Figure 4: Screenshot with annotations from our supplementary "results.html" which provides an interface to easily switch back-and-forth between different results.

to select "the best looking images". The results of this are shown in Table 2, some entries are N/A since the authors did not provide the necessary dereflection results in their papers or project websites. Even though our method only takes 2 frames as input, it outperforms [2, 3] and is on par with the recently published [4]. Our dual-view dereflection approach is only outperformed by [4, 10] on their own test images.

5. Dataset Quality

Existing training datasets for reflection removal could unfortunately not have been used to supervise a dual-view model, we thus had to make our own dataset which we will make publicly available. One may argue that our improved results for reflection removal may primarily stem from hav-

ing better training data than existing work. However, our rendered dual-view training data is subject to a significant domain gap and our rendering pipeline only models glass as mirrors with alpha transparency. As for our training data based on transformed real-world images, it follows the image acquisition approach of [9, 12] and the image formation model of [3, 8, 11]. Our new dual-view training dataset thus provides no benefit for supervising single-image dereflection models. This is exemplified by [12] performing better on the real-world test set of our main user study than our single-view ablation (our results are preferred 87% of the time over [12], whereas our results are preferred 92% of the time over our single-view ablation). However, on our single-view ablation performs better than [12] on our rendered test set (quantitatively and qualitatively as shown in Table 3 of our main paper). This indicates that there is a domain gap between our synthetic dataset and real-world footage.

6. Qualitative Comparison Tool

Please see the provided "results.html" to visually compare our proposed approach to several competing approaches for reflection removal. We provide this interactive interface, of which an annotated screenshot is shown in Figure 4, to help the reviewers to better assess the quality of our results.

References

[1] Damien Fourure, Rémi Emonet, Élisa Fromont, Damien Muselet, Alain Trémeau, and Christian Wolf. Residual Conv-Deconv Grid Network for Semantic Segmentation. In *British Machine Vision Conference*, 2017. 1

[2] Xiaojie Guo, Xiaochun Cao, and Yi Ma. Robust Separation of Reflection From Multiple Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[3] Yu Li and Michael S. Brown. Exploiting Reflection Change for Automatic Reflection Removal. In *IEEE International Conference on Computer Vision*, 2013. 2, 3

[4] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to See Through Obstructions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[5] Simon Niklaus and Feng Liu. Context-Aware Synthesis for Video Frame Interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[6] Suji Shimizu, Toshiharu Kondo, Takashi Kohashi, M Tsurata, and Teruyoshi Komuro. A New Algorithm for Exposure Control Based on Fuzzy Logic for Video Cameras. *IEEE Transactions on Consumer Electronics*, 38(3):617–623, 1992. 1

[7] Sudipta N. Sinha, Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski. Image-Based Rendering for Scenes With Reflections. *ACM Transactions on Graphics*, 31(4):100:1–100:10, 2012. 2

[8] Richard Szeliski, Shai Avidan, and P. Anandan. Layer Extraction From Multiple Images Containing Reflections and Transparency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000. 3

[9] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single Image Reflection Removal Beyond Linearity. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[10] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A Computational Approach for Obstruction-Free Photography. *ACM Transactions on Graphics*, 34(4):79:1–79:11, 2015. 2

[11] Jiaolong Yang, Hongdong Li, Yuchao Dai, and Robby T. Tan. Robust Optical Flow Estimation of Double-Layer Images Under Transparency or Reflection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[12] Xuaner Cecilia Zhang, Ren Ng, and Qifeng Chen. Single Image Reflection Separation With Perceptual Losses. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3