4-2021

Learned Dual-View Reflection Removal

Simon Niklaus
Adobe Research

Xuaner Cecilia Zhang
UC Berkeley

Jonathan T. Barron
Google Research

Neal Wadhwa
Google Research

Rahul Garg
Google Research

See next page for additional authors

Citation Details

Niklaus, S., Zhang, X. C., Barron, J. T., Wadhwa, N., Garg, R., Liu, F., & Xue, T. (2021). Learned Dual-View Reflection Removal. *Institute of Electrical and Electronics Engineers (IEEE).* https://doi.org/10.1109/wacv48630.2021.00376

Simon Niklaus, Xuaner Cecilia Zhang, Jonathan T. Barron, Neal Wadhwa, Rahul Garg, Feng Liu, and Tianfan Xue

Learned Dual-View Reflection Removal

Simon Niklaus
Adobe Research

Xuaner (Cecilia) Zhang
UC Berkeley

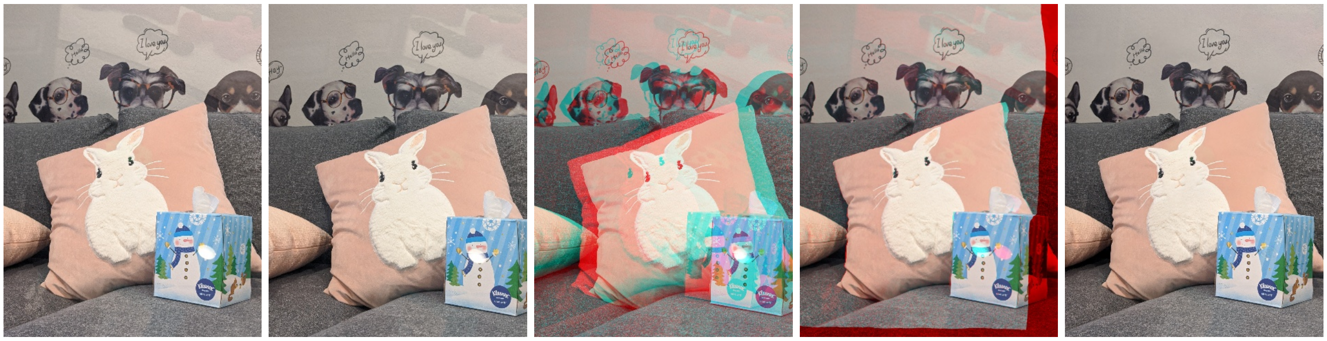
Jonathan T. Barron
Google Research

Neal Wadhwa
Google Research

Rahul Garg
Google Research

Feng Liu
Portland State University

Tianfan Xue
Google Research



(a) View 1 (b) View 2 (c) Anaglyph (d) Aligned (e) Ours

Figure 1: Stereo pairs (a, b) were imaged through glass and exhibit undesired reflections. The transmitted and reflective images are subject to parallax that is difficult to separate as shown in the anaglyph (c). Our reflection-invariant flow aligns the two views with respect to the transmitted image, causing all remaining parallax (in the reflection on the tissue box, for example) to be due to reflections as shown in anaglyph (d). Our synthesis network exploits this parallax to remove reflections (e).

Abstract

Traditional reflection removal algorithms either use a single image as input, which suffers from intrinsic ambiguities, or use multiple images from a moving camera, which is inconvenient for users. We instead propose a learning-based dereflection algorithm that uses stereo images as input. This is an effective trade-off between the two extremes: the parallax between two views provides cues to remove reflections, and two views are easy to capture due to the adoption of stereo cameras in smartphones. Our model consists of a learning-based reflection-invariant flow model for dual-view registration, and a learned synthesis model for combining aligned image pairs. Because no dataset for dual-view reflection removal exists, we render a synthetic dataset of dual-views with and without reflections for use in training. Our evaluation on an additional real-world dataset of stereo pairs shows that our algorithm outperforms existing single-image and multi-image dereflection approaches.

1. Introduction

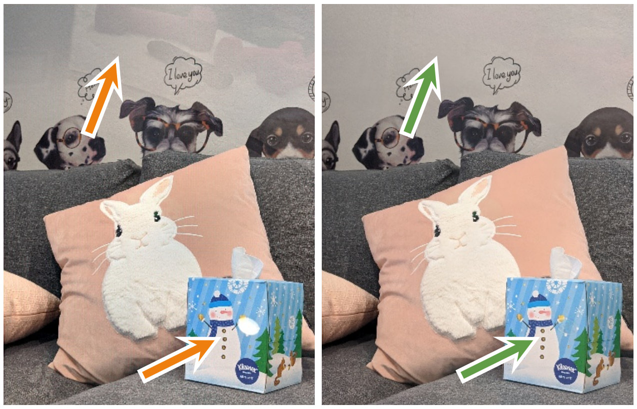
Of the billions of pictures taken every year, a significant portion are taken through a reflective surface such as a glass window of a car or a glass case in a museum. This presents a problem for the photographer, as glass reflects some of the incident light from the same side as the photographer back towards the camera, corrupting the captured images with reflected image content. Formally, the captured image I is the sum of the image being transmitted through the glass T and the image of the light being reflected by the glass R:

$$I[x, y, c] = T[x, y, c] + R[x, y, c]. \tag{1}$$

The task of reflection removal is estimating the image T from an input image I. A solution to this problem has significant value, as it would greatly broaden the variety of circumstances in which photography can occur.

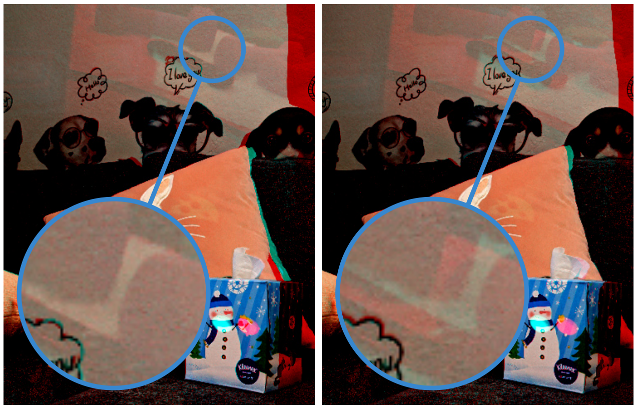
Equation 1 shows the core difficulty of single-image reflection removal: the problem is inherently underconstrained, as we have six unknowns at each pixel but only three observations. Most single-image techniques for reflection removal try to mitigate this problem by using image priors to disambiguate between reflection and transmission. Despite significant progress, most algorithms still cannot cleanly separate them. In fact, even humans may have difficulty when just given a single image. For example, it is difficult to tell whether the white spot next to the snowman in Figure 1(a) is a reflection or not without having a second perspective.

Work primarily done while Simon and Xuaner were interns at Google.



(a) Ablation - Using One View (b) Ours - Using Two Views

Figure 2: Comparison of a single-view ablation (a) to our proposed dual-view reflection removal (b). Reasoning jointly about both views allows our proposed approach to handle challenging scenes like this one. In comparison, the single-view ablation fails to remove all present reflections due to the underconstrained nature of the single-image setting.



(a) Traditional Optical Flow (b) Reflection-Invariant Flow

Figure 3: Aligned stereo anaglyphs by warping I_2 to I_1 with traditional optical flow (a), and our reflection-invariant optical flow (b). Contrast adjusted for visualization. Traditional flow aligns all image content, minimizing the parallax in both transmission and reflection. With our reflection-invariant optical flow, all remaining parallax is in the reflection.

The ambiguity of the single-image case led to the development of multi-image techniques. Figure 1(a) and 1(b) show two views of a scene in which the camera translates slightly. Because the reflective and transmissive layers do not have the same distance from the camera, the scene content of the reflective layer moves differently from the transmissive layer when switching between the two views as shown in Figure 1(c). This parallax can help to disambiguate between reflection and transmission, thereby simplifying the task of recovering the constituent images. For this reason, practical systems for reflection removal rely on acquiring many images or entire videos of the same subject under different viewpoints [24, 39]. However, this setup is burdensome as it requires users to manually move their camera while capturing many images, and it assumes a static scene.

This points to a fundamental tension between single-image and multi-image techniques. We explore a compromising solution in which we take as input two views of the same scene produced by a stereo camera (Figure 2). Though binocular stereo is not new, smartphones are adopting camera arrays, thereby increasing the practicality of algorithms designed for stereo images. This presents an opportunity for high-quality dual-view dereflection that is as convenient as any single-image technique, requiring just a single button press and being capable of capturing non-static scenes.

Still, it is not trivial to extend existing single- or multi-image dereflection algorithms to dual-view input. Most multi-image algorithms [39, 43] use hand-tuned heuristics based on motion parallax and require at least 3 to 5 frames as input, as two views are often not enough to make this problem well-posed. And most single-image dereflection algorithms [8, 16, 38, 45] are trained on images with synthetic reflections, a strategy which does not generalize to dual-view input due to the need for realistic motion parallax.

To address these issues, we combine merits of both approaches and propose a learned approach that utilizes motion parallax. We first align the two input images using the motion of only the transmissive layer. Ignoring reflective content during registration produces aligned images where the transmissive layer is static while the reflection “moves” across aligned views, reducing the transmission-reflection separation problem to one of simply distinguishing between static and moving edges, as shown in Figure 3(b). Unlike traditional flow approaches, which align both transmissive and reflective image content as shown in Figure 3(a), we explicitly train an optical flow network to be invariant to reflections. After performing this reflection-invariant alignment, we supervise a image synthesis network to recover the transmission from the transmission-aligned views.

While this framework is conceptually simple, training such a model requires difficult-to-acquire dual-view imagery that is subject to reflections. It is even more difficult to obtain such data with accurate ground truth optical flow of the transmissive layer. As such, we resort to employing computer graphics and render virtual environments to create such a dataset. We also collect a real-world dual-view dataset with ground truth transmission for evaluation purposes, and show that our approach generalizes well to this data.

2. Related Work

The task of reflection removal is a narrow sub-problem of the classical problem of inferring a complete model of the physical world that generated an observed image [4], which has been extensively studied throughout the history of

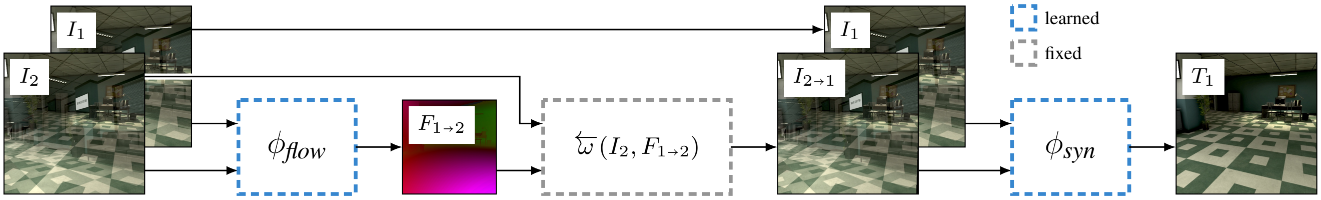


Figure 4: Our dual-view reflection removal. Given images I_1 and I_2, our reflection-invariant optical flow network ϕ_{flow} estimates the motion $F_{1\to 2}$ between the unknown transmissive layers of the inputs, which is then used to warp I_2 towards I_1 to "undo" that motion. Our synthesis network ϕ_{syn} can then use these aligned images to leverage the parallax between the reflective layers (and the lack thereof between the transmissive layers) to synthesize I_1's transmissive layer T_1.

computer vision. Reflection removal is similar in nature to other blind signal separation problems in computer vision, such as disentangling reflectance and shading [3] or separating haze from transmitted scene content [12]. Due to the ill-posed nature of reflection removal, many past works used additional information to constrain the problem. A common strategy is to use multiple images captured from different viewpoints as input, taking advantage of how transmitted content is constant across images while the reflective content changes [11, 22, 24, 39]. These approaches require significant labor from the photographer, and also assume a static scene. Another approach is to use multiple images from the same view but with different polarization [18, 30], which leverages the relationship between the angle of incidence of light on the reflecting surface and its polarization. Though effective, these techniques require a static scene and the rather exotic ability to modify a camera's polarization.

Automatic single-image reflection removal techniques are an attractive alternative to multi-image solutions [35]. Prior to the rise of deep learning, single-image reflection techniques would usually impose beliefs about the natural world or the appearance of reflected images, and then recover the transmittance and reflectance that best satisfy those priors. These approaches require the manual construction of regularizers on edges or relative smoothness [20, 23, 31, 42], then solving an expensive and/or non-convex optimization problem. With deep learning, the focus shifted towards training a network to map from the input image to the transmission [8, 21, 36, 40, 45]. Though effective, these techniques depend critically on the quality of training data.

Our work addresses an unexplored approach that lies between single-image and multi-image cases. By combining the information present in stereo imagery with the effectiveness of a neural network trained on vast amounts of synthetic data, our approach produces higher-quality output than single-image approaches while requiring none of the labor or difficulty of multi-image approaches.

Stereo cameras are closely related to dual-pixel sensors, wherein a single camera has a sensor with "split" pixels, thereby allowing it to produce limited light fields [10, 34]. Dual-pixel reflection removal has been explored with promising results [28], but it is unclear how such a technique might generalize to stereo. First, the dual-pixel disparity is only significant in cameras with large apertures, like DSLRs but not smartphones. When using a DSLR though, reflections are out of focus and are heavily blurred which in itself already provides important cues. Second, due to the interplay between focus distance and dual-pixel images, one can simply threshold the dual-pixel disparity to separate reflection edges from transmitted content as done in [28]. Such a universal threshold does unfortunately not exist for stereo images.

3. Method

Given images I_1 and I_2 captured from two different viewpoints, our goal is to estimate T_1, an image that contains only the transmissive content of I_1. We have found that a single network is unable to synthesize T_1 from I_1 and I_2 directly, presumably due to the difficulty of simultaneously aligning and combining these images. We hence decompose this task into: reflection-invariant motion estimation, warping to account for transmission parallax, and transmission synthesis. We recover the optical flow $F_{1\to 2}$ between the transmissive layers of I_1 and I_2 using a network ϕ_{flow} as

$$F_{1\to 2} = \phi_{flow}\left(I_1, I_2\right) \qquad (2)$$

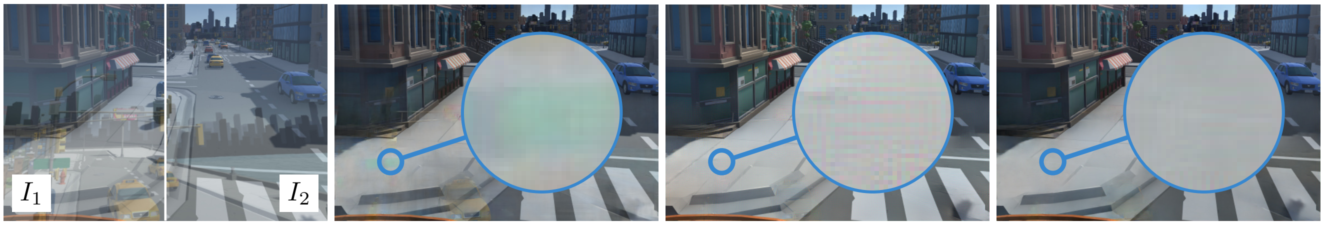
This step depends critically on ϕ_{flow} being trained to be invariant to reflection, as we describe in Section 3.1. We then use this optical flow to account for the inter-frame transmission motion via differentiable sampling [13]. Specifically, we use backward warping $\overleftarrow{\omega}$ and warp I_2 to I_1 according to the estimated optical flow $F_{1\to 2}$ to generate $I_{2\to 1}$ as

$$I_{2\to 1} = \overleftarrow{\omega}\left(I_2, F_{1\to 2}\right), \qquad (3)$$

Because our optical flow is reflection-invariant, I_2 is warped such that only its transmissive content matches that of I_1. This allows us to apply a synthesis model that takes as input the image of interest I_1 and its warped counterpart $I_{2\to 1}$, and estimates the first image's transmissive layer T_1 as

$$T_1 = \phi_{syn}\left(I_1, I_{2\to 1}\right). \qquad (4)$$

Combining these Equations 2–4 gives our complete reflection removal pipeline, which we also visually summarize in Figure 4, where ϕ_{flow} and ϕ_{syn} are neural networks.



(a) Input (b) $\mathcal{L}_1$ (c) $\mathcal{L}_F$ (d) $\mathcal{L}_{\text{LPIPS}}$

Figure 5: Training with ℓ_1 distance led to low-frequency artifacts (b), and using squared distance between VGG features led to checkerboard artifacts (c). We hence train our synthesis model using LPIPS, which produces good results (d).

3.1. Reflection-Invariant Optical Flow

Most learning-based optical flow models assume that each pixel has a single motion and train on datasets where this assumption holds [5, 6]. However, in the presence of reflections, each pixel can have two valid motions: that of the transmission and that of the reflection. Applying learned flow models trained on existing datasets to images containing reflections produces motion estimates that are a compromise between the two true underlying motions, causing them to work poorly for our dereflection task. We hence train a reflection-invariant flow estimation network using our own synthetic dataset which we introduce in Section 3.3. We do so by adopting the architecture of PWC-Net [32] and supervising it for $1.5 \cdot 10^6$ iterations with 8 samples per batch and a learning rate of 10^{-4} using TensorFlow's default Adam [17] optimizer on our new synthetic dataset.

Thanks to our new dataset, our flow model is largely invariant to reflections. In comparison, a model supervised on a reflection-free version of our dataset is subject to a significant drop in its flow prediction accuracy once reflections are introduced (Section 4.1). This reflection-invariant flow estimate is critical to make our dereflection approach work and an ablation of our pipeline with a regular optical flow network fails to produce convincing results (Section 4.2).

3.2. Dual-View Transmission Synthesis

Given the first view I_1 and the aligned second view $I_{2\to 1}$, we utilize a neural network to synthesize the desired transmissive layer T_1 of I_1. In doing so, the aligned view $I_{2\to 1}$ provides important cues which allow the synthesis network to produce high-quality results despite the presence of significant reflections. Because our optical flow network produces motion estimates that are invariant to reflections, transmissive image content in these warped images is aligned but reflective content is not aligned as long as there is motion parallax between them. This reduces the burden on the synthesis model, as even a pixel-wise minimum of two images should produce good results, as demonstrated in [33].

We use a GridNet [9] with the modifications from Niklaus *et al.* [26] for our synthesis network, using five rows and four columns where the first two columns perform downsampling and the last two columns perform upsampling. GridNets are a generalization of U-Nets [29], which are often used for image synthesis tasks. In essence, GridNets allow information within the network to be processed along multiple streams at different resolutions, which enables them to learn how to combine features across different scales.

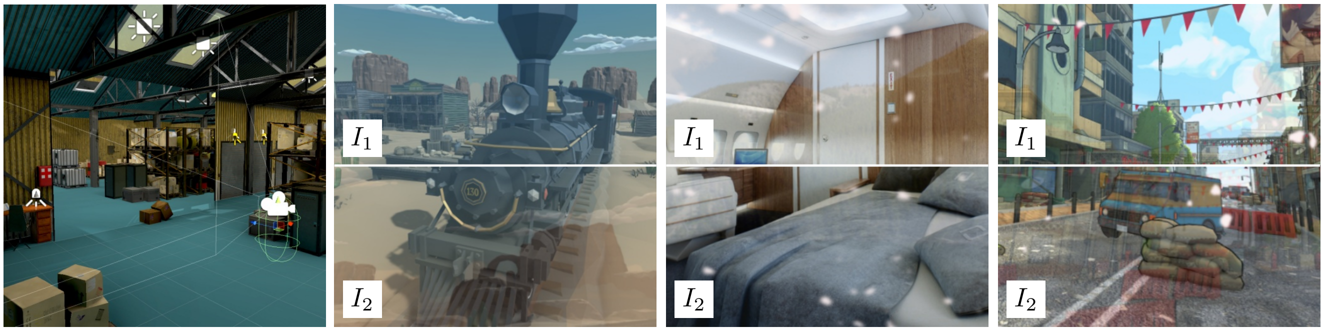
We supervise this synthesis model on our dual-view dataset, which we describe in Section 3.3. Instead of directly using the ground truth optical flow to warp I_2 towards I_1, we use the prediction of our reflection-invariant optical flow network. This forces the trained synthesis model to be more robust with respect to misaligned transmissions that may be introduced by erroneous optical flow estimates.

We analyzed several possible loss functions to supervise our synthesis model. The simplest of which is the ℓ_1 distance between the predicted transmission layer and ground truth. However, a synthesis model supervised with just $\mathcal{L}_1$ is prone to low-frequency artifacts as shown in Figure 5(b). We additionally explored a loss based on the squared distance between VGG features [15], which some recent dereflection algorithms have used successfully [45]. However, we noticed subtle checkerboard artifacts when supervising our synthesis model on this $\mathcal{L}_F$ as shown in Figure 5(c) (even when using bilinear upsampling instead of transposed convolutions [27]). We thus used the LPIPS metric [44], which linearly weights feature activations using a channel-wise vector w as

$$\mathcal{L}_{\text{LPIPS}} = \sum_{\ell} \left\| w^{\ell} \odot \left(\Phi^{\ell}\left(T_1^{pred}\right) - \Phi^{\ell}\left(T_1^{gt}\right)\right)\right\|_2^2 . \quad (5)$$

Specifically, we use version "0.1" of this metric, using AlexNet [19] to compute feature activations, and where the weights w have been linearly calibrated to minimize the perceptual difference in accordance with a user study [44]. Our synthesis model trained using $\mathcal{L}_{\text{LPIPS}}$ is able to produce pleasant results that are not subject to checkerboard artifacts, as shown in Figure 5(d). This perceptual loss serves a similar purpose as adversarial losses, which have also been an effective mean for the task of reflection removal [45].

We train our proposed dual-view transmission synthesis model using TensorFlow's default Adam [17] optimizer with a learning rate of $5 \cdot 10^{-5}$, which took a total of 1.5 million iterations with 4 samples per batch to fully converge.



(a) Unity Editor (b) Rendered Views (c) Warped Images (d) Warped Renders

Figure 6: Our training dataset consists of three different types of images: 60% are fully-rendered images generated using the Unity engine (a) and consist of scenes with complex geometry (b), 30% are real images that lack ground-truth geometry and have instead been warped using random homographies to generate the second view (c), and 10% are warped rendered images to make sure that the model does not "cheat" (d). Note that because (b) is fully rendered, its reflective layer originates from the same domain as the transmissive layer (both are mountains), while the two layers in (c) may have different sources.

3.3. Dual-View Training Data

Existing learning-based methods for dereflection combine pairs of images to synthesize training data [8, 45]. This approach works well for monocular approaches, but it does not generalize to our dual-view approach. After all, whatever reflection we add to a stereo pair should be geometrically consistent across the two views which requires difficult-to-acquire depth maps. Furthermore, training our reflection-invariant flow network requires ground truth optical flow between the transmissive layers of the two views. However, acquiring ground truth flow is a challenging problem with previous work having exploited hidden fluorescent textures, computer graphics, and high frame-rate videos [2, 5, 14].

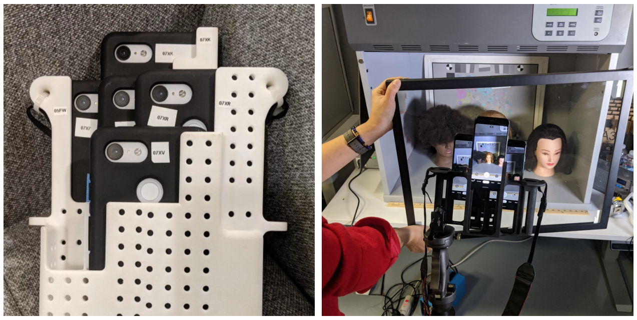
For these reasons, we rely on computer graphics to synthesize our training data. We acquired 20 virtual environments from professional artists, 17 of which are used for training and 3 of which are used for evaluation. These environments vary greatly, and include indoor scenes, cityscapes, and naturalistic scenes. We render them with Unity, which allowed us to collect arbitrary views together with a ground-truth inter-frame optical flow. Views are generated by pre-recording camera paths through the scene, from which we sample camera locations for I_1. We generate I_2 by randomly shifting the position of I_1 by up to 0.5 meters and randomly rotating the camera by up to 10 degrees. To model reflections, we create a translucent mirror that is placed in front of the two cameras. We uniformly sample the mirror's alpha blending factor $\alpha \sim \mathcal{U}(0.6, 0.9)$, and apply a Gaussian blur with a random $\sigma \sim \mathcal{U}(0.0, 0.1)$ to the reflective image to mimic depth of field. We then alpha-blend the transmissive and reflective images to get the rendered output for I_1 and I_2.

Training only on synthetic data may result in poor performance on real-world data, due to a significant gap between the two domains [25]. To address this, we augment our synthetic data with additional training data that has been generated using real-world images. We first randomly sample two images and blend them to get the input for one view, and apply two homography transforms to the two images independently to synthesize the image in the other view. This basically assumes that the transmissive and reflective layers are on independent planes. Although this over-simplifies the geometry of the real world compared with our fully-rendered data, it helps the network to better fit to the statistics of real-world images. We collected 7000 images with a Creative Commons license for this purpose and manually selected those with pleasant visual aesthetics, which yielded a subset of 1000 images in total. As shown Figure 6(c), this data is closer to real world imagery but it lacks real motion parallax. Warping image I_2 to image I_1 according to the transmission flow is hence free from disocclusions. This is not the only unrealistic aspect of this approach though, since reflections may not originate form the same scene like as in the picture of a hotel room that exhibits reflections of a mountain.

To make sure that our model does not "cheat" by identifying which images are real and taking advantage of our simple proxy geometry, we also applied the same homography-based image formation model that was used for our real-world data to our rendered data, as shown in Figure 6(d).

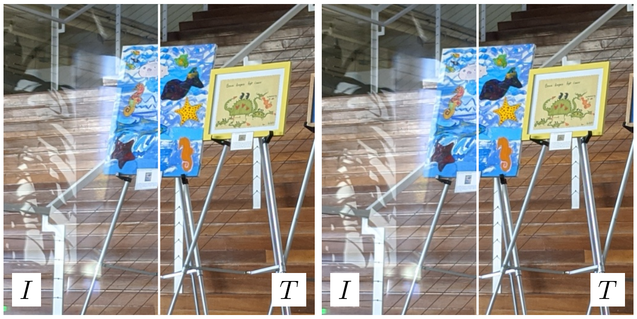
Lastly, many reflections in the real world stem from light sources which yield saturated bright spots in the image. To model this, we augment the reflective layer with a mask of bright spots obtained from binarized fractal noise: we compute the fractal noise from Perlin noise at 4 octaves with a persistence uniformly drawn from $\rho \sim \mathcal{U}(0.3, 1.0)$ before binarizing the mask based on a threshold of 1. To avoid unnatural discontinuities, we further apply a Gaussian blur with $\sigma \sim \mathcal{U}(1, 5)$ to this binary mask. Examples of such saturated bright spots are shown in Figure 6(c) and 6(d).

When using this training dataset, we randomly sample 60% of the batches from our rendered data, 30% from our warped images, and 10% from our warped renderings.



(a) Camera (b) Setup

Figure 7: A picture of our custom-built camera rig consisting of five synchronized Google Pixel phones (a) as well as a schematic reenactment of the data capturing setup (b).



(a) Without Alignment (b) With Alignment

Figure 8: The images in our dataset and their respective transmissions are misaligned due to refraction (a), as can be seen at the stairs. We align them to account for this (b).

4. Experiments

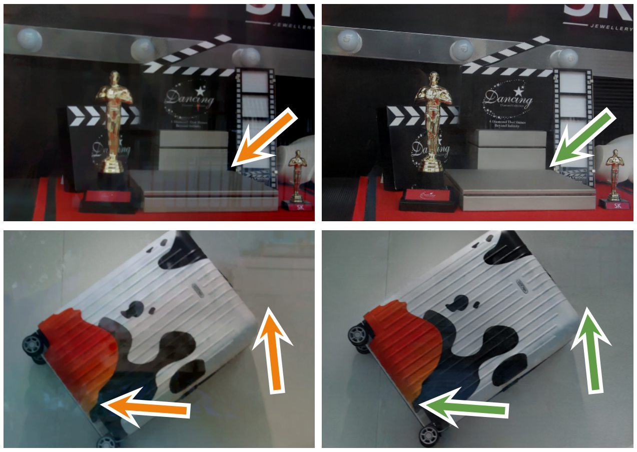
We evaluate on rendered and real-world images.

Rendered test set: To build a rendered test set, we used 3 virtual worlds that are not used in training and rendered 60 different samples. We also recorded the corresponding ground truth transmission image without reflection and the ground truth optical flow between the transmission layers.

Real-world test set: To build a real-world test set, we use a camera rig of five phones as shown in Figure 7 and synchronize them using [1]. To test that our approach works for different stereo configurations, we always use the center camera as the reference view and one of the other four cameras as the second view. For each of the 20 scenes we captured, we obtained the transmission and between 2 and 4 sets of images with reflections by placing different types of glass in front of the camera. As discussed in [39], the transmission shifts between the image capturing with the glass and without the glass due to refractions unless the glass is infinitely thin. Therefore, we register the image captured through glass to the ground truth transmission (image captured without glass) using an affine transform calculated by [7]. An example of this alignment is shown in Figure 8.

	rendered test w/o refl.				rendered test w/ refl.			
	EPE mean	EPE median	ABS mean	ABS median	EPE mean	EPE median	ABS mean	ABS median
Zeros	24.90	22.88	24.54	24.00	24.90	22.88	24.54	24.00
Oracle	0.0	0.0	3.13	2.88	0.0	0.0	3.13	2.88
Train w/o refl.	1.14	0.84	4.02	3.56	4.52	2.67	6.10	5.56
Train w/ refl.	1.53	1.05	4.23	3.67	2.39	1.26	4.68	4.06

Table 1: Flow accuracy on our rendered test set. We trained two versions of our flow network, one using our rendered test set w/ reflections and one w/o reflections. We also report the accuracy of zero and ground truth motion as bounds.



(a) Robust Flow [41] (b) Ours

Figure 9: Comparisons with [41], a robust optical flow via classic energy minimization, on examples from their paper.

4.1. Reflection-Invariant Optical Flow

Metrics: Following optical flow literature [2], we use two metrics to evaluate flow accuracy: 1) the end-point error (EPE) between the estimated flow and the true flow, and 2) the absolute difference (ABS) between the first frame and the second frame warped to the first frame using the estimated flow. For the ABS metric, as we only calculate the motion of the transmission layer, we only warp the ground truth transmission layer without reflection even though the motion was estimated from the input images with reflection. We also mask out the occluded pixels based on the true transmission optical flow when calculating the ABS metric.

Results: Table 1 shows the quantitative results. To better understand the scale of EPE and ABS, we also report these metrics for zero flow (all pixels are static) and ground truth transmission flow ("Oracle"). Note that because of lighting changes between left and right views, the ABS error of the ground truth flow is not zero. When evaluating on input with reflection, the flow network trained with reflection is more robust than the one trained without reflection, with 47% less mean EPE error and 23% less mean ABS error. We analyze

		rendered test set			real-world test set		
	images used	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Zhang-like	1	23.92	0.872	0.137	22.39	0.742	0.124
Mono	1	26.31	0.928	0.068	22.35	0.752	0.110
Concat	2	25.81	0.927	0.069	22.10	0.752	0.111
Regular Flow	2	17.61	0.827	0.099	17.73	0.684	0.128
Ours	2	26.60	0.938	0.058	22.82	0.765	0.104

Table 2: Results from our ablation study, showing the importance of GridNet and reflection-invariant optical flow.

	rendered test set				real-world test set			
	quantitative			users	quantitative			users
	PSNR ↑	SSIM ↑	LPIPS ↓	prefer ours	PSNR ↑	SSIM ↑	LPIPS ↓	prefer ours
Input	23.38	0.887	0.155	99%	22.25	0.761	0.114	95%
Zhang *et al.*	22.21	0.811	0.217	99%	21.47	0.725	0.172	87%
Wen *et al.*	22.34	0.856	0.185	100%	21.56	0.744	0.142	94%
Li & Brown	22.00	0.794	0.243	100%	20.49	0.671	0.227	98%
Ours - Mono	26.31	0.928	0.068	94%	22.35	0.752	0.110	92%
Ours	26.60	0.938	0.058	–	22.82	0.765	0.104	–

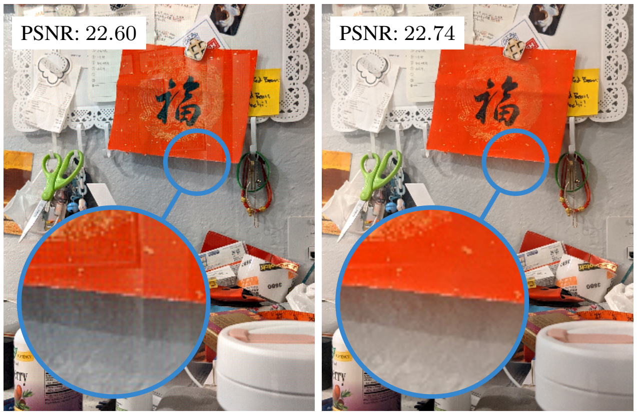
Table 3: Quantitative evaluation of the recovered transmission image, together with the results from a user study with responses from 20 participants across 9 rendered test images and 20 real test images. Users were asked to compare our dual-view result to one of five baselines. We report the percentage of times that users preferred our method.

the effect of this difference in the context of our reflection removal pipeline in the ablation study in Section 4.2.

Related: Optical flow estimation on layered compound images has previously been studied by Yang *et al*. [41], who proposed a solution based on classic energy minimization. We were unable to use this technique as a baseline on our benchmark, as the implementation provided by the authors does not allow for arbitrary images to be processed (it requires some external optical flow estimate as input). We hence compare to this technique by instead applying our dereflection pipeline to the example images used by [41]. As can be seen in Figure 9, our proposed approach produces significantly improved reflection removal results.

4.2. Dual-View Transmission Synthesis

Metrics: To quantitatively evaluate the quality of reflection removal, we use three evaluation metrics: PSNR, the hand-designed similarity metric SSIM proposed by Wang *et al*. [37], and the learned similarity metric LPIPS proposed by Zhang *et al*. [44]. Because the transmission coefficient of glass is less than 1.0, the transmission captured through the glass is dimmer than the image captured without glass. As a result, there is an unknown scaling factor between the estimated transmission and the ground truth. To make our evaluation invariant to this unknown scaling factor, we first scale the estimated transmission by searching for the gain s and bias b that minimize $\|s \cdot T_1^{\text{pred}} + b - T_1^{\text{gt}}\|_2$, before computing the error metrics using the scaled estimate.

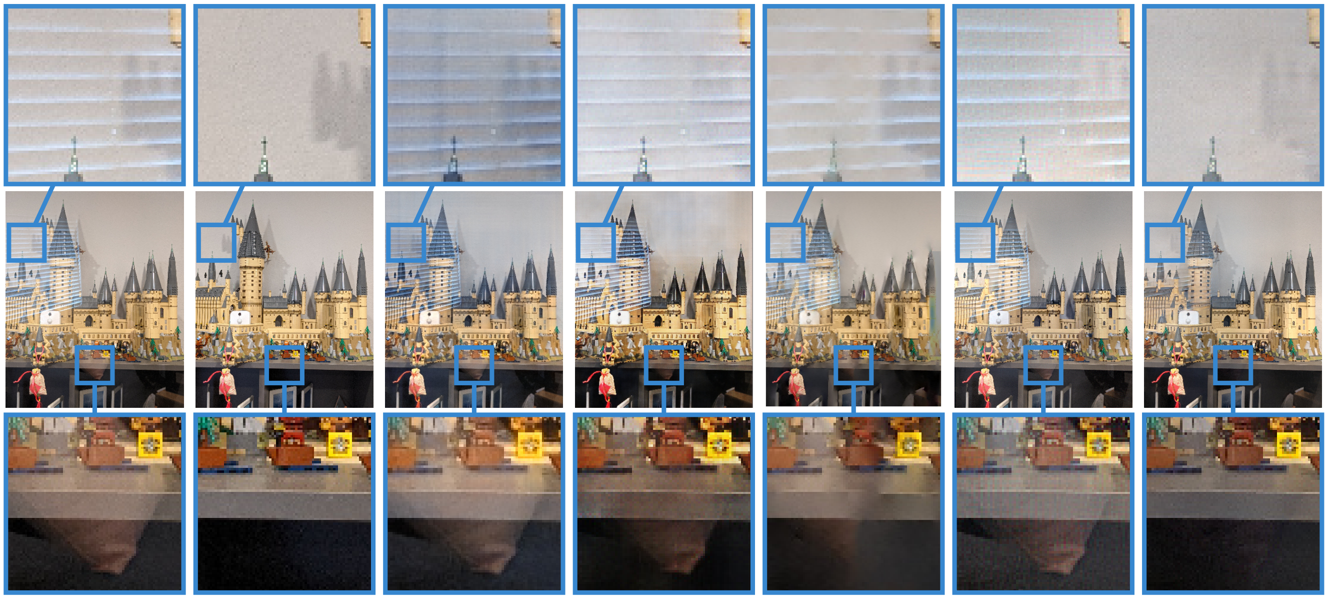


(a) Ours - Mono (b) Ours

Figure 10: A result from our mono baseline (a) and our approach (b). They have a comparable PSNR, yet 19 out of 20 participants in a user study preferred the result of (b).

Ablation: We analyzed different components of our proposed network composition in an ablation study and tried four variations: 1) "Zhang-like", i.e., training the model from Zhang *et al*. [45] on our dataset, 2) "Mono", by only using a single input, 3) "Concat", by concatenating the input images without explicitly aligning them first, and 4) "Regular Flow", by replacing the flow network with the one trained on images without reflection. Table 2 shows the quantitative results. "Mono" outperforms "Zhang-like", which shows that the GridNet network architecture is well suited to this task. Also, our network with reflection invariant flow outperforms both "Concat" and "Regular Flow". This exemplifies the importance of reflection-invariant alignment.

Quantitative: The quantitative comparison of the recovered transmission image is shown in Table 3, it includes comparisons to four baseline algorithms: two single-frame reflection removal algorithms by Zhang *et al*. [45] and Wen *et al*. [38], one multi-frame algorithm by Li and Brown [22], and a single-image ablation of our approach ("Ours - Mono"). Our proposed dual-view approach outperforms all baselines on all metrics, demonstrating the effectiveness of our method. However, using the input image itself as a baseline already shows surprisingly good results, especially on the real-world test dataset. This raises the question of whether or not traditional quality metrics are suitable for evaluating reflection removal. This is exemplified by Figure 10, which shows example results with similar PSNR but a strong preference by human examiners for one over the other. We thus subsequently further compare the results though a user study.



Input Truth Zhang *et al*. [45] Wen *et al*. [38] Li & Brown [22] Ours - Mono Ours

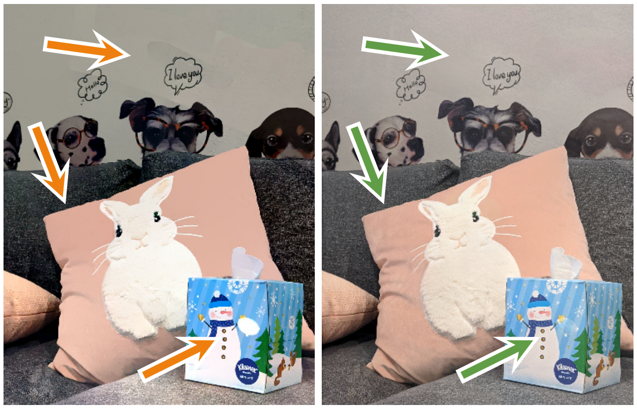
Figure 11: Qualitative comparison. Please see the supplementary material for a tool-supported visual comparison.

User study: We conducted an A/B user study with 20 participants that were not related to this project, including 2 professional photographers, to further evaluate our results. We chose subsets for each test set to keep the number of comparisons for each participant below 200. For our rendered test set, we chose 3 challenging samples from each virtual test world resulting in 9 images. For our real-world test set, we chose the center and right cameras from the first capture in each set, resulting in 20 images. We asked each participant to select "the best looking images". The results of this are included in Table 3. Overall, our approach is preferred over the baselines in the vast majority of cases.

Qualitative: We show a representative example result in Figure 11, which shows that our proposed dual-view approach can better remove challenging reflections in our test data. Please also consider the supplementary material for a comparison tool which includes many more examples.

4.3. Dual-Pixel Reflection Removal

Recently, Punnappurath *et al*. [28] proposed a dual-pixel reflection removal technique. Dual-pixel images superficially resemble stereo pairs in that they both capture two perspectives of a scene. However, this dual-pixel technique performs poorly when applied to our stereo data: it achieved a PSNR/SSIM/LPIPS score of 17.82/0.774/0.230 on our rendered test set and 14.52/0.567/0.350 on our real-world test set (examples shown in Figure 12). This is consistent with recent work on dual-pixel imagery for depth estimation [10], which has shown that dual-pixel footage is sufficiently different from stereo in terms of photometric properties that it benefits from being treated as a distinct problem domain.



(a) Dual Pixels [28] (b) Ours

Figure 12: On our stereo data, the recent dual-pixel technique [28] flattens textures and does not catch all reflections.

5. Conclusion

In this paper, we presented a new learning-based dual-view reflection removal approach. Unlike the traditional reflection removal techniques, which either take a single frame or multiple frames as input, we proposed to use dual-view inputs, which yields a nice trade-off between the convenience of capturing and the resulting quality. To train this learned dual-view dereflection approach, we created a new dual-view dataset by rendering realistic virtual environments. We also designed a new composite network consisting of a reflection-invariant optical flow estimation network and a dual-view transmission synthesis network. We have shown promising experimental results on both synthetic and real images with challenging reflections, outperforming previous work.

References

[1] Sameer Ansari, Neal Wadhwa, Rahul Garg, and Jiawen Chen. Wireless Software Synchronization of Multiple Distributed Cameras. In *IEEE International Conference on Computational Photography*, 2019. 6

[2] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision*, 92(1):1–31, 2011. 5, 6

[3] Jonathan T. Barron and Jitendra Malik. Shape, Illumination, and Reflectance From Shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. 3

[4] Harry Barrow and Martin Tenenbaum. Recovering Intrinsic Scene Characteristics. *Computer Vision Systems*, 2(3-26):2, 1978. 2

[5] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *European Conference on Computer Vision*, 2012. 4, 5

[6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow With Convolutional Networks. In *IEEE International Conference on Computer Vision*, 2015. 4

[7] Georgios D. Evangelidis and Emmanouil Z. Psarakis. Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008. 6

[8] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David P. Wipf. A Generic Deep Architecture for Single Image Reflection Removal and Image Smoothing. In *IEEE International Conference on Computer Vision*, 2017. 2, 3, 5

[9] 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. 4

[10] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T. Barron. Learning Single Camera Depth Estimation Using Dual-Pixels. In *IEEE International Conference on Computer Vision*, 2019. 3, 8

[11] Byeong-Ju Han and Jae-Young Sim. Reflection Removal Using Low-Rank Matrix Completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[12] Kaiming He, Jian Sun, and Xiaoou Tang. Single Image Haze Removal Using Dark Channel Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. 3

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*, 2015. 3

[14] Joel Janai, Fatma Güney, Jonas Wulff, Michael J. Black, and Andreas Geiger. Slow Flow: Exploiting High-Speed Cameras for Accurate and Diverse Optical Flow Reference Data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*, 2016. 4

[16] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single Image Reflection Removal With Physically-Based Training Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv/1412.6980*, 2014. 4

[18] Naejin Kong, Yu-Wing Tai, and Joseph S. Shin. A Physically-Based Approach to Reflection Separation: From Physical Modeling to Constrained Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):209–221, 2014. 3

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification With Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 2012. 4

[20] Anat Levin, Assaf Zomet, and Yair Weiss. Separating Reflections From a Single Image Using Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 3

[21] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E. Hopcroft. Single Image Reflection Removal Through Cascaded Refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[22] Yu Li and Michael S. Brown. Exploiting Reflection Change for Automatic Reflection Removal. In *IEEE International Conference on Computer Vision*, 2013. 3, 7, 8

[23] Yu Li and Michael S. Brown. Single Image Layer Separation Using Relative Smoothness. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3

[24] Ce Liu, Michael Rubinstein, Mike Krainin, and Bill Freeman. PhotoScan: Taking Glare-Free Pictures of Pictures. Technical report, 2017. 2, 3

[25] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018. 5

[26] Simon Niklaus and Feng Liu. Context-Aware Synthesis for Video Frame Interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[27] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and Checkerboard Artifacts. Technical report, 2016. 4

[28] Abhijith Punnappurath and Michael S. Brown. Reflection Removal Using a Dual-Pixel Sensor. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 8

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv/1505.04597*, 2015. 4

[30] Yoav Y. Schechner, Joseph Shamir, and Nahum Kiryati. Polarization-Based Decorrelation of Transparent Layers: The Inclination Angle of an Invisible Surface. In *IEEE International Conference on Computer Vision*, 1999. 3

[31] Yi-Chang Shih, Dilip Krishnan, Frédo Durand, and William T. Freeman. Reflection Removal Using Ghosting Cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[32] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[33] 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. 4

[34] Neal Wadhwa, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T. Barron, Yael Pritch, and Marc Levoy. Synthetic Depth-Of-Field With a Single-Camera Mobile Phone. *ACM Transactions on Graphics*, 37(4):64:1–64:13, 2018. 3

[35] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Benchmarking Single-Image Reflection Removal Algorithms. In *IEEE International Conference on Computer Vision*, 2017. 3

[36] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. CRRN: Multi-Scale Guided Concurrent Reflection Removal Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[37] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[38] 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. 2, 7, 8

[39] 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, 3, 6

[40] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing Deeply and Bidirectionally: A Deep Learning Approach for Single Image Reflection Removal. In *European Conference on Computer Vision*, 2018. 3

[41] 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. 6, 7

[42] Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, and Weiyu Xu. Fast Single Image Reflection Suppression via Convex Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[43] Zhefei Yu, Houqiang Li, Zhangyang Wang, Zeng Hu, and Chang Wen Chen. Multi-Level Video Frame Interpolation: Exploiting the Interaction Among Different Levels. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(7):1235–1248, 2013. 2

[44] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4, 7

[45] 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. 2, 3, 4, 5, 7, 8