Revisiting Single Image Reflection Removal In the Wild

Yurui Zhu [1, 2, *], Xueyang Fu [1, ✉], Peng-Tao Jiang [2],
Hao Zhang [2], Qibin Sun [1], Jinwei Chen [2], Zheng-Jun Zha [1], Bo Li [2, ✉]
[1] University of Science and Technology of China [2] vivo Mobile Communication Co., Ltd
zyr@mail.ustc.edu.cn, xyfu@ustc.edu.cn, libra@vivo.com

Abstract

This research focuses on the issue of single-image reflection removal (SIRR) in real-world conditions, examining it from two angles: the collection pipeline of real reflection pairs and the perception of real reflection locations. We devise an advanced reflection collection pipeline that is highly adaptable to a wide range of real-world reflection scenarios and incurs reduced costs in collecting large-scale aligned reflection pairs. In the process, we develop a large-scale, high-quality reflection dataset named Reflection Removal in the Wild (RRW). RRW contains over 14,950 high-resolution real-world reflection pairs, a dataset forty-five times larger than its predecessors. Regarding perception of reflection locations, we identify that numerous virtual reflection objects visible in reflection images are not present in the corresponding ground-truth images. This observation, drawn from the aligned pairs, leads us to conceive the Maximum Reflection Filter (MaxRF). The MaxRF could accurately and explicitly characterize reflection locations from pairs of images. Building upon this, we design a reflection location-aware cascaded framework, specifically tailored for SIRR. Powered by these innovative techniques, our solution achieves superior performance than current leading methods across multiple real-world benchmarks. Codes and datasets are available at here.

1. Introduction

In photographic environments involving reflective materials, such as glass, the inadvertent emergence of reflections is a common challenge. These underside reflections not only diminish the aesthetic quality of the captured images but also impede the accuracy of follow-up computer vision tasks [24, 30, 37]. Consequently, devising effective reflection removal algorithms is important and meaningful.

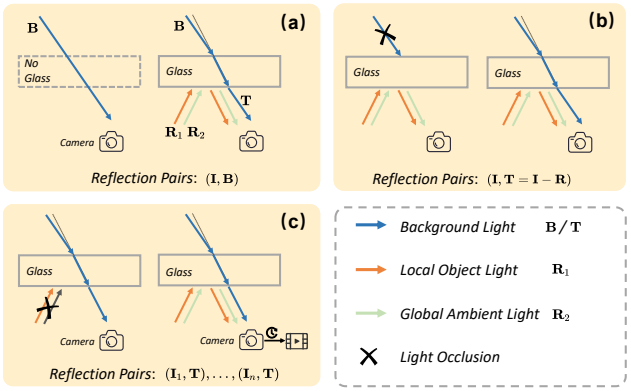

Figure 1. Simplified illustrations of existing pipelines for collecting real reflection pairs (**I**: reflection image). (a) The pipeline from [19, 39, 53] may lead to misalignment between **B** and **T** due to glass refraction. (b) The pipeline from [15, 16] is limited to RAW data format collection, and the obtained **T** may contain reflection remnant artifacts. (c) Our approach avoids issues related to glass refraction and artifacts, and does not impose data format constraints. Furthermore, our video-based capture system reduce the difficulty and effort involved in large-scale data acquisition.

Deep learning-based SIRR methods have recently exhibited encouraging results. It's well acknowledged that an ample supply of high-quality data is essential for these data-driven methods. Accordingly, a range of datasets has been developed to support research in SIRR. Nevertheless, our thorough examination of the collection pipelines corresponding to these datasets reveals consistently overlooked issues in Figure 1. For instance, pioneering studies [19, 46, 53] acquire reflection-free images by manually removing the glass. However, this technique invariably induces spatial pixel misalignment due to the refraction triggered by the glass. Lei *et al*. [15, 16] exploit the linear reflection formation in raw space to extract the transmission layer with the subtraction operation. However, as depicted in Figure 2, we noticed that their method might leave minor reflection remnants in the corresponding transmission images. Furthermore, it's worth noting that the effort and

⋆ : This work was done during his internship at vivo Mobile Communication Co., Ltd.
✉: Corresponding authors.

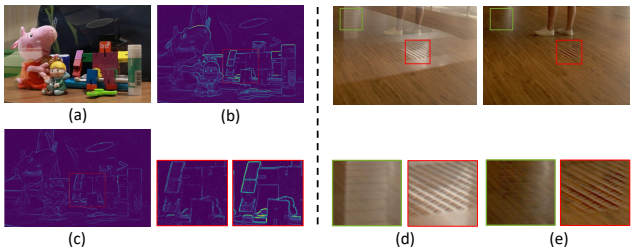

Figure 2. Analysis of collected images from the previous collection pipelines. (a) Reflection image from the acquisition pipeline [19, 39, 53]; (b) Gradient of (a); (c) Gradient difference map between the reflection pair, where the 'double edge' in the gradient difference map indicates misalignment due to glass refraction. (d) and (e) are the reflection pair from the acquisition pipeline [15], where minor reflection remnants could be found in the corresponding transmission images. (**Best viewed on screen.**)

time necessary to expand the dataset size based on previous pipelines can be quite costly. Therefore, the costs associated with prior data collection pipelines, along with issues like misalignment or artifacts, cause a shortage of high-quality, large-scale real pairs for training deep models.

In this study, we revisit the reflection physical formulation process. This subsequently leads to the development of an innovative pipeline for the collection of real reflection pairs. To be specific, the essence of reflection disturbances lies in the fact that the reflected lights bounce off the surface of the reflective material and are captured by camera devices, thereby disturbing the clarity of the transmission layer. Hence, we could directly obtain reflection-free images by obstructing the reflective light, thereby enabling the acquisition of pairs of images with and without reflections. As illustrated in Figure 1, our pipeline captures the transmission layer without the removal of glasses, utilizing non-transmissive and non-reflective black cloth to block reflection lights. Subsequently, the light-blocking materials are removed, facilitating the capture of images with reflection distortions. It is noteworthy that the whole process is implemented in video mode, reducing the difficulty and efforts involved in large-scale data acquisition. During the collection phase, we could dynamically manipulate reflective contents, thereby enhancing the diversity of reflection distortions. Evidently, compared to other pipelines, our pipeline enables us to more easily acquire the aligned and extensive real reflection pairs for training deep networks.

Upon further revisiting the reflection imaging process, we identify that the constituents of reflections can be categorized into two parts: global reflections induced by ambient light and local reflections caused by specific objects. The latter, local reflections, are typically more challenging, prompting many innovative solutions [4, 40] to address them. For example, Wan *et al*. [40] assume that the gradients of reflections are generally small and obtain the reflection-dominated regions via the threshold algorithm. Indeed, such an approach is not suitable for scenarios with strong reflections. Dong *et al*. [4] utilize the linear composition loss to implicitly infer the reflection confidence maps. Yet, their linear assumption often falls short in describing the complex real-world scenes. Such location cues are known to effectively mitigate reflection disturbances, as demonstrated by studies [4, 39, 40]. However, these techniques rely on prior assumptions to indirectly obtain the desired results, which often come with certain limitations.

In this paper, we highlight that if the collection pipelines enable collecting aligned reflection pairs, then directly utilizing these aligned paired images can effectively characterize the locations of these local reflections. Specifically, we observe the fact that the reflection layer encompasses textures of many virtual objects, which are absent in the corresponding ground-truth images. Building upon this observation, we devise the maximum reflection filter (MaxRF) that could explicitly present the reflection locations. Therefore, this paper proposes a divide-and-conquer framework tailored for SIRR, including reflection detection and removal. Within our frameworks, we distinguish the local reflection regions based on representations derived from MaxRF. Subsequently, these location cues are integrated into the second stage of our framework, significantly enhancing the performance of removing reflections. Finally, experimental results also demonstrate the effectiveness and the superiority of our proposed solution. Contributions of this paper could be summarized as:

- We propose a new pipeline for the collection of real reflection pairs. notable for its adaptability to a wide range of reflection scenarios and its independence from data format constraints. This pipeline also offers a more cost-effective manner of acquiring reflection datasets.
- We present a large-scale high-quality paired reflection dataset, Reflection Removal in the Wild (RRW). To the best knowledge, RRW is the largest paired reflection dataset, comprising 14952 pairs of high-resolution images captured across diverse real-world reflection scenes.
- We propose the maximum reflection filter (MaxRF), which enables obtaining the explicit location representation to characterize reflection regions.
- We develop a cascaded network for SIRR, which involves the reflection detection and removal network, *i.e*., first learning to estimate reflection locations with MaxRF and then removing reflections with location guidance. Comprehensive experiments indicate the superiority of the proposed innovations.

2. Related work

Over recent decades, numerous innovative methods have been proposed to tackle the issue of image reflection removal. Some approaches usually require additional inputs, such as multi-frames [1, 20, 28], polarization [15, 27, 34],

and flash-only prior [14, 17]. In this paper, our primary focus lies in the field of single-image reflection removal.

Traditional methods. Early methods [2, 7, 14, 18, 21, 35, 48, 54] utilize various image priors to eliminate reflection degradation. For example, Li *et al*. [21] devise a relative smoothness prior, postulating that the reflection contents are intrinsically blurry, consequently penalizing these larger gradients. Shih*et al*. [35] propose to automatically suppress reflection by leveraging "ghosting" cues from double reflections on thick glasses and employing a Gaussian Mixture Model for regularization. In [18], the user annotations are used to guide layer separation between the transmission and reflection layers. Nikolaos *et al*. [2] impose the laplacian data fidelity term and gradient sparsity as optimization objectives. Despite producing decent results, traditional methods often rely on assumptions and tend to have slower processing speeds.

Learning-based methods. With the development of deep-learning techniques, learning-based SIRR methods [10, 23, 26, 29, 41–43, 55] also achieve great performance gains in the diverse reflection scenes and dominate this field. Concretely, CEILNet [5] adopts a two-stage network approach, which first estimates the edge map and subsequently reconstructs the transmission layer. ERRNet [46] attempts to leverage high-level contextual features to mitigate uncertainty in these regions with prominent reflections and introduce the misaligned real-world pair images. Yu *et al*. [22] and BDNet [50] both incorporate the reflection layer to guide the restoration of the transmission layer. Song *et al*. [4] further investigates the robustness of SIRR networks against adversarial attacks. In addition, previous methods [9, 10] have explored the complementary mechanism and developed the dual-stream frameworks to achieve reflection separation. Moreover, LANet [4] proposes a location-aware solution with a recurrent network to remove reflections in single images, improving results by emphasizing strong reflection boundaries with Laplacian features. Unlike LANet using the implicit manner to estimate the reflections, our proposed method perceives the reflection location with the explicit representation.

Various methods are also devoted to addressing the insufficiency of real-world training data, categorizing them into three primary avenues. The first direction is to model reflections that are more consistent with real-world scenarios. For example, Kim*et al*. [12] utilize the physically-based rendering to generate various reflection image pairs for training. Wen *et al*. [47] propose to exploit the non-linearity capability of deep neural networks to simulate the real-world physical reflection imaging process. However, these synthetic reflection images are still far from the real-world physical formulation, which may bring the performance drop under real reflection scenes. The second approach utilizes unsupervised and weakly supervised algorithms to alleviate the demand for large-scale paired training data. [31] proposed an unsupervised SIRR method by optimizing the two deep network parameters to separate the target image into exclusive transmission and reflection layers. However, compared to supervised learning techniques, the performance of these methods [31, 32] still exhibits considerable room for further advancement.

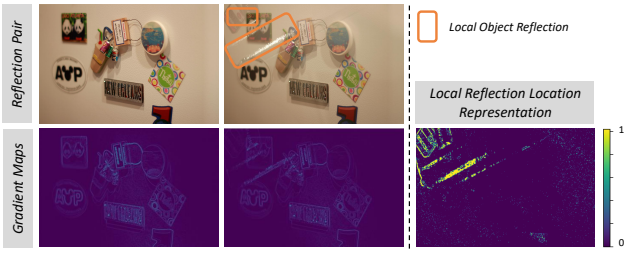

Figure 3. Visualization of the local reflection location presentation via our proposed Maximum Reflection Filter (MaxRF). First column: non-reflection images. Second column: reflection images The global ambient reflection causes attenuation of the color and contrast, subsequently diminishing the gradient intensity of the original transmission contents. Hence, applying MaxRF predominantly could well highlight local reflection locations. (**Best viewed on screen.**)

The third direction involves collecting paired reflections from the real world. For example, methods [19, 39, 53] capture pair samples where images with the manual arrangement of glass represent reflection images, while those without the glass serve as reflection-free images. Moreover, [15, 16] observe the linear formulation of the reflection imaging process is held on raw space, and they attempt to obtain the transmission layer by subtraction operation. However, in this paper, compared with previous ones, our pipeline could avoid the pixel misalignment introduced by glass refraction and does not make assumptions regarding the data format. Moreover, our pipeline is more applicable for diverse reflective scenes and enables the acquisition of large-scale reflection image pairs at a lower cost.

3. Revisit Reflection Physical Formulation

We revisit the physical formulation underlying the occurrence of reflections, specifically using reflection scenarios involving glasses reflection as examples. Within the reflection-contaminated image, the physical lights are a mixture of reflection and transmission lights. In this section, we further analyze these two components and clarify ambiguities of SIRR identified in previous studies.

For the former, we first define the light originating from behind the glasses as the background **B**, and the light that passes through the glass as the transmission **T**. As cameras shoot these objects behind the glasses, the background light undergoes refraction and absorption [54], sub-

sequently transforming into the transmission light. This results in a distinct difference between the background **B** and the transmissions **T**, as depicted in Figure 1. Notably, the inherent refraction property of glass indicates that earlier data acquisition pipelines[19, 46, 53] inherently led to pixel misalignment, making it impossible to obtain perfectly aligned image pairs especially with thick glass. Moreover, prevailing studies [49, 52] suggest that the light transmitted through glass substantially surpasses the absorbed part, indicating the absorption effects can be roughly neglected. Therefore, according to the above analysis, we argue that SIRR should focus on removing the illumination disturbances from the camera's side and obtaining clean transmission **T**, which is also consistent with [15].

Besides, reflections commonly manifest when a camera captures illumination rays reflected off surfaces within its field of view. We noted that physical aspects of reflections can be broadly categorized into two distinct components. As shown in the first row of Figure 3, the first involves global reflection resulting from ambient light, causing attenuation of the color and contrast in the captured images. The second pertains to the virtual contents formed after the reflection from objects on the camera side, which typically occupies a part of the whole image. Such local reflections typically result in occlusions or overlapping with the transmission contents. Previous methods attempt to identify the spatial location of these local reflections either through the indirect paradigm, *e.g.*, gradient prior [40] or via implicit constraints [4, 19]. In contrast, we propose the maximum reflection filter(MaxRF) to directly acquire reflection locations from reflection pairs. Subsequently, we employ neural networks to learn and distinguish such local reflection regions, as discussed in Sec. 4.1.

4. Method

4.1. Explicit Reflection Location Perception

The virtual image formed by object reflections often occupies only a portion of the image, and it is typically the more challenging reflection component to handle. In this paper, we devise an explicit representation to characterize these reflection locations.

Due to the global ambient light reflection, the difference between the reflection pairs cannot directly obtain the local reflection location information. However, in fact, regardless of the strength of these local reflections, it often results in the presence of texture details in the reflection-contaminated image that is absent in the transmission layer. Hence, we propose the Maximum Reflection Filter (MaxRF) to identify these reflection regions. To elaborate, MaxRF involves two key steps. Firstly, we apply the Sobel operator [6] to compute gradient maps for both the reflection and reflection-free images. Secondly, we perform

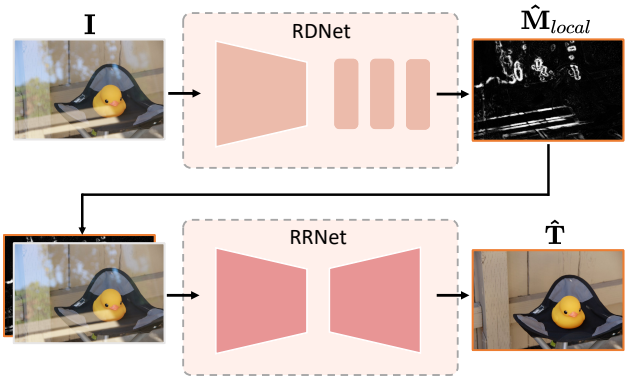

Figure 4. Simplified illustration of our proposed framework, including the RDNet (Reflection Detection Network) and RRNet (Reflection Removal Network).

maximum comparisons in the corresponding gradient domain. It's worth noting that due to the influence of global ambient light reflection, the intensity of gradients associated with the original transmissions is also reduced compared to their initial strength. As a result, when we apply the MaxRF, what remains mainly are the gradients indicative of the local reflection contents. Finally, we could employ the reflection image pairs to explicitly obtain the local reflection locations, denoted as $\mathbf{M}_{local}$, which could be expressed as:

$$
\begin{aligned}
G_I, G_T &= Grad(\mathbf{I}), Grad(\mathbf{T}), \\
\mathbf{M}_{local}^{(i,j)} &= \begin{cases} 1 & \text{if } G_I^{(i,j)} > G_T^{(i,j)}, \\ 0 & \text{otherwise}, \end{cases}
\end{aligned} \tag{1}
$$

where **I** and **T** indicate the reflection-contaminated image and transmission image, respectively; $Grad(\cdot)$ indicates the Sobel gradient operator; and $\mathbf{M}_{local}^{(i,j)} = 1$ indicates the presence of the local reflection at the spatial position (i, j). Moreover, we also visualize the local reflection location presentation via our proposed (MaxRF) in Figure 3.

4.2. Network Architecture

In this section, we further elaborate the proposed cascaded network architecture tailored for SIRR. As shown in figure 4, this structure comprises two primary components: the reflection detection network (RDNet) and the reflection removal network (RRNet). For RDNet, we employ the pre-trained network backbone [38] in conjunction with several residual blocks [8] and interpolation operation to estimate $\mathbf{M}_{local}$. For RRNet, we utilize the estimated $\hat{\mathbf{M}}_{local}$ to guide the subsequent reflection removal. Concretely, we adopt the widely used restoration backbone [3] as our baseline structure for reflection removal. The estimated $\hat{\mathbf{M}}_{local}$ and the input reflection image **I** are concatenated and subsequently fed into RRNet. The whole process could be for-

mulated as:

$$\hat{\mathbf{M}}_{local} = \text{RDNet}(\mathbf{I}), \tag{2}$$

$$\hat{\mathbf{T}} = \text{RRNet}(Concat([\mathbf{I}, \hat{\mathbf{M}}_{local}])). \tag{3}$$

Based on Eqn. 1, we can directly utilize the reflection image pair to derive the explicit representation for local reflection regions. Consequently, we adopt the supervised learning manner to estimate $\hat{\mathbf{M}}_{local}$. The loss function for RDNet is defined as follows:

$$\mathcal{L}_{DNet} = \left\| \mathbf{M}_{local} - \hat{\mathbf{M}}_{local} \right\|_1 + \gamma_1 * TVLoss(\hat{\mathbf{M}}_{local}), \tag{4}$$

where γ_1 represents the balancing weight for the $TVLoss$ [33], adopted to smooth the estimated results and mitigate artifacts. Moreover, for RRNet, we employ the content loss and perceptual loss as defined in [10, 19, 53], which are expressed as:

$$\mathcal{L}_{RNet} = \left\| \mathbf{T} - \hat{\mathbf{T}} \right\|_1 + \gamma_2 * \left\| VGG(\mathbf{T}) - VGG(\hat{\mathbf{T}}) \right\|_1, \tag{5}$$

where γ_2 indicates the balanced weight; $VGG(\cdot)$ indicates hierarchical features extracted by the four layers $conv1\_2$, $conv2\_2$, $conv3\_2$, and $conv4\_2$ of the VGG19 [36].

4.3. Dataset Collection Pipeline

As illustrated in Figure 1, we compare our proposed pipeline with those previously presented in [15, 19, 53]. Figure 1(a) employs images captured both before and after the manual placement of glass, which facilitates capturing both reflection-containing images and their background counterparts. Nonetheless, their approach fails to consider the refraction effect and color of the glasses. Consequently, when the glasses are thick or colored, it can lead to noticeable pixel misalignment or color distortion. Meanwhile, following this pipeline, it is impractical to capture reflection image pairs by involving the common glass scenes in daily life, restricting its broad applicability. Moreover, such misaligned pairs can even pose challenges for network training [15]. On the other hand, both Figure 1(b) and ours recognize the refraction effect of glass. However, the second pipeline is based on the observation of the linear physical composition held on raw data. They initially capture the reflection-contaminated image $\mathbf{I}$ and the reflection layer $\mathbf{R}$ in raw format, and then produce the transmission layer through $\mathbf{T} = \mathbf{I} - \mathbf{R}$. However, empirically, we notice that the obtained transmission layers frequently retain subtle reflection remnants.

In contrast to the previous, our pipeline is applicable to a wide range of reflection scenarios (*e.g.*, various glass scenes) and relaxes the requirement of the data format. Specifically, we commence the collection process by employing a black velvet cloth to block reflective lights on the camera side, ensuring the acquisition of the clean transmission image, denoted as $\mathbf{T}$. Once the cloth is removed, we can then acquire images with reflection disturbances. More importantly, our pipeline is implemented in video mode. During this phase, we further actively modulate the contents (*e.g.*,) on the reflective side to diversify the reflective scenes. These modulation operations include blocking reflective lights and adjusting or introducing reflective objects, among others. Therefore, our proposed pipeline also facilitates the scaling up of real-world reflection training datasets at a lower cost.

Table 1. Summary of existing real reflection datasets.

Dataset	Year	Usage	Pairs Number	Average Resolution
SIR^2	2017	Test	454	540 × 400
$Real$	2018	Train / Test	89/20	1152 × 930
$Nature$	2020	Train / Test	200/20	598 × 398
RRW(Ours)	2023	Train	14952	2580 × 1460

Our proposed dataset is primarily captured using two camera devices: the Apple iPhone 13 and a Digital Single-Lens Reflex (DSLR) Canon EOS 200DII. In total, we have collected video clips from nearly 150 unique scenes and sampled 14,952 pairs of reflection images. To ensure alignment between reflection image pairs ($\mathbf{I}$ and $\mathbf{T}$), the tripod and the remote control shutter are used for camera stabilization. Moreover, we provide comparisons with other datasets in Table 1 and show some reflection examples in Figure 5.

5. Experiments

5.1. Implementation details

Our framework is implemented with PyTorch platform on a PC with NVIDIA GeForce GTX 1080Ti. At the training phase, the network is trained by Adam [13] optimizer with an initial learning rate of 0.0006, which changes based on Cosine Annealing scheme [25].The two sub-networks are jointly trained for about 60 hours with four 1080Ti GPUs. The batch size is set to four, and the 320 × 320 patches are randomly cropped from the image at each training iteration. The hyperparameters in Eqns. 4 5 are empirically set as $\gamma_1 = 0.00005$, $\gamma_2 = 0.02$.

5.2. Dataset and Evaluation Metrics

During the training phase, we enhance the training dataset by integrating data used in previous methods [4, 9, 10, 19], with additional data collected RRW dataset, providing a more comprehensive training database. For the testing dataset, following previous methods, we evaluate the performance of our model by applying one pre-trained reflection removal model across three real-world reflection benchmarks: $Real$, $Nature$, and SIR^2. These three benchmarks, developed through different works, comprehensively cover a variety of real-world reflection scenarios,

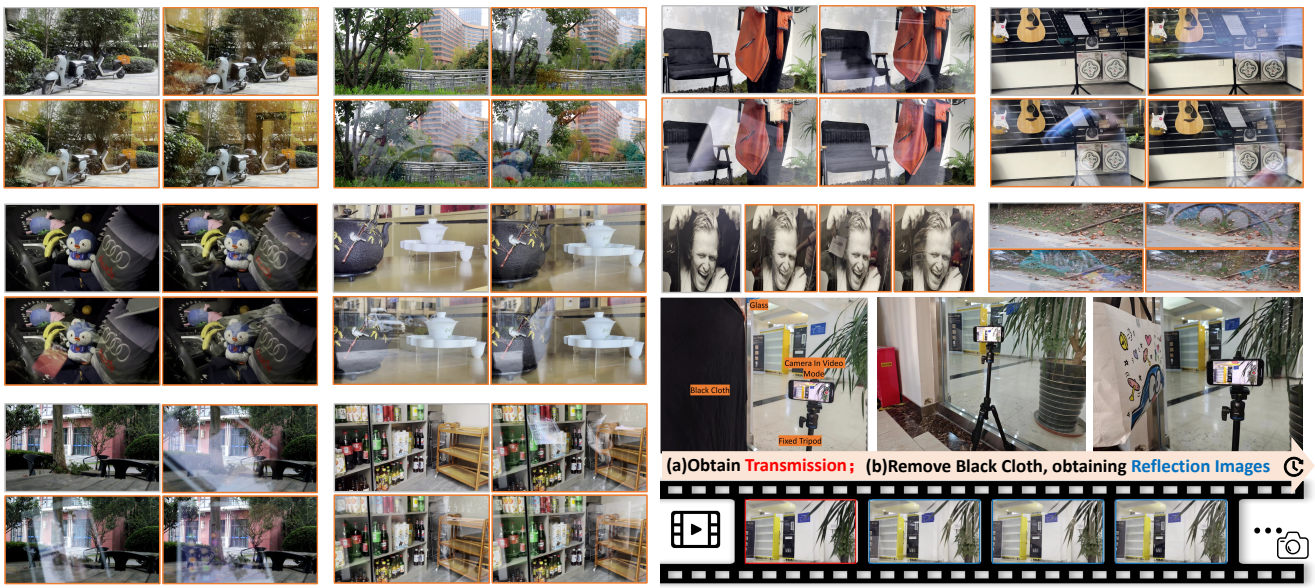

Figure 5. Some reflection pairs of the proposed RRW dataset and the data collection process with our proposed pipeline. Our pipeline is a video mode capture system, where each non-reflection image (gray boxes) corresponds to multiple real reflection images (orange boxes), encompassing a variety of reflective surfaces such as building glasses, car glass windows, display glass, framing glass, and self-prepared glass. The diversity demonstrates the pipeline's applicability across various reflection scenarios.

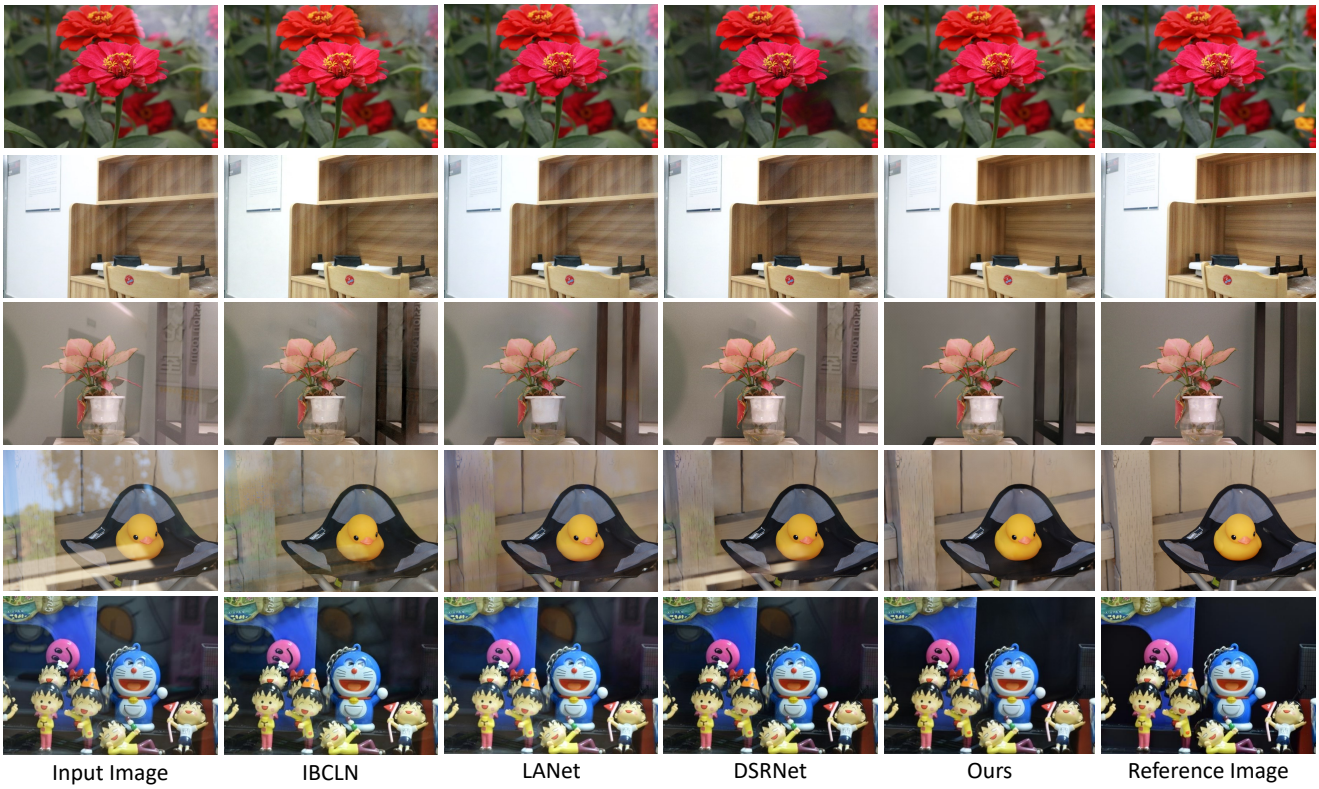

Figure 6. Visual comparisons between our method and previous methods. Unless otherwise specified, all reflection images in this paper are from real-world reflection scenes. More visual results are available in our supplemental material.

Table 2. Quantitative comparisons on the real reflection benchmarks. The best results are in **bold**, and the second-best results are underlined.

Methods	Venue	$Nature$ (20)		$Real$(20)		SIR^2(454)		$Average$(494)	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Input Image	-	20.44	0.785	18.96	0.733	22.76	0.885	22.51	0.884
BDN [50]	ECCV 2018	18.83	0.738	18.64	0.726	21.61	0.854	21.50	0.844
FRS [51]	CVPR 2019	20.01	0.756	18.63	0.719	22.23	0.867	21.99	0.867
Zhang *et al*. [11]	CVPR 2018	22.31	0.804	20.16	0.767	23.07	0.869	22.92	0.862
ERRNet [46]	CVPR 2019	22.57	0.807	20.67	0.781	22.97	0.885	22.85	0.877
RMNet [47]	CVPR 2019	21.08	0.730	19.93	0.718	21.66	0.843	21.57	0.834
Kim *et al*. [12]	CVPR 2020	20.10	0.759	20.22	0.735	23.57	0.877	23.30	0.886
IBCLN [19]	CVPR 2020	23.90	0.787	21.42	0.769	24.05	0.888	23.94	0.878
YTMT [9]	NerIPS 2021	20.69	0.777	22.94	0.815	23.57	0.889	23.43	0.882
LANet [4]	ICCV 2021	23.51	0.810	23.40	**0.826**	23.04	0.898	23.07	0.891
PNACR [44]	ACM MM 2023	23.93	0.807	22.57	0.806	24.14	0.894	24.06	0.888
DSRNet [10]	ICCV 2023	21.24	0.789	22.32	0.806	24.91	0.902	24.65	0.893
Ours	-	**25.96**	**0.843**	**23.82**	0.817	**25.45**	**0.910**	**25.40**	**0.904**

which are widely used for showcasing the performance of models in real-world reflection scenarios. However, among these three datasets $Real$, $Nature$, and SIR^2, the first two include both training and test datasets, while the latter inherently serves as the test dataset.Besides, we employ the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [45] as the evaluation metrics. These are calculated in the RGB color space.

5.3. Comparison to State-of-the-arts

To evaluate the reflection removal performance, we compare our proposed method with 11 SIRR methods, including BDN [50], FRS [51], Zhang *et al*. [11], ERRNet [46], RMNet [47], Kim *et al*. [12], IBCLN [19], YTMT [9], LANet [4], PNACR [44], and DSRNet [10]. For fair comparisons, we directly employ the pre-trained weights publically provided by their authors. Following [10, 46], the comparison experiments are also performed under the same settings, such as using the same reflection inputs and the same performance evaluation codes. Additionally, for methods [11, 46], additional finetuning is implemented. This was due to the fact that these methods were developed before the introduction of the $Nature$ dataset, and as such, their initial training dataset does not encompass the training data of the $Nature$ dataset. Note that we do not perform finetuning on RMNet [47], due to their method relays on additional alpha blending masks from SynNet [47]. Hence, apart from the differences in the design of the algorithm compared to other methods, our solution further utilizes the collected real-world dataset RRW, to construct a more comprehensive training dataset.

The quantitative comparison results are reported in Table 2. Our proposed method obviously achieves the best PSNR scores across all real benchmarks, which effectively demonstrates superior performance against the previous SOTA methods. Specifically, our approach is 2.03dB higher than the second-best method [44] on the PNSR metric. The last column reports that our method also achieves the best average PSNR and SSIM scores. This verifies the powerful generalization capacity on the various real reflection cases.

Figure 6 further provides visual comparisons of reflection removal results from three SOTA methods and ours. These real reflection inputs all from $Real$, $Nature$, and SIR^2 dataset. For example, in the third row of input images, global ambient reflections and local object reflections are both present. IBCLN [19] and LANet [4], although significantly mitigating the global color and contrast distortions induced by global reflections, still exhibit residual local reflections. DSRNet [10], while successful in diminishing certain local reflections, exhibits a noticeable disparity in color and contrast compared to the reference image. We observe in the fourth row that other methods, when dealing with stronger reflections, tend to introduce artifacts such as color biases. In contrast, our approach effectively eliminates both types of reflections and retains high-frequency transmission details in the estimated results.

5.4. Ablation study

Visualization of the reflection location. Figure 7 shows the estimated location maps of reflection regions, including the previous method (LANet [4]) and ours. In their method, LANet employs the linear composition loss to implicitly deduce the reflection confidence maps. Since the distribution of reflection distortions is often uneven, their learned confidence maps could be used to simply depict the locations of the reflection regions. In contrast, We directly utilize the reflection representation obtained from MaxRF as the objective to optimize the reflection detection network (RDNet). Compared to LANet, which represents reflection locations based on weighted composition weights, our approach offers a more explicit manner to directly characterize local reflections. As a result, the visualization results in Figure 7 show that our method can estimate the locations of local reflections more accurately and clearly.

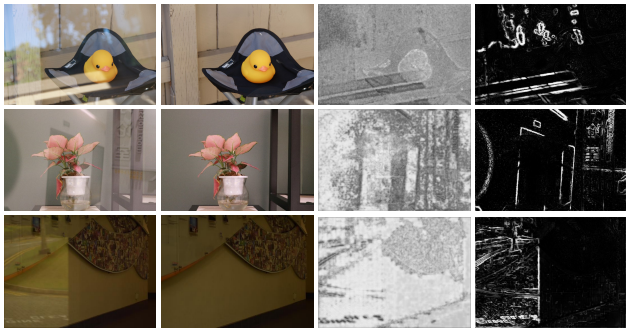

Figure 7. Estimated results of the local reflection locations. (a) Reflection images; (b) reference images; (c) reflection confidence maps of LANet [4]; (d) the estimated location maps of Ours.

Table 3. Ablation study of the components of our framework. Extra Data indicates that we incorporate the RRW as the additional dataset during the training phase. The PSNR↑ metric is used.

Model Configurations			*Nature*	*Real*	SIR^2
RDNet	RRNet	Extra Data			
✓	✓		25.37	22.65	24.93
	✓	✓	25.49	23.38	24.82
✓	✓	✓	25.96	23.82	25.45

Table 4. Extension of our proposed innovations to pioneer works.

Model Configurations	*Nature*	*Real*	SIR^2
Zhang *et al.* [53]	22.31	20.16	23.07
+RDNet	23.02	20.76	23.46
+RDNet + Extra Data	23.57	21.40	23.79
ERRNet [46]	22.57	20.67	22.97
+RDNet	23.64	21.15	23.57
+RDNet + Extra Data	24.15	21.73	23.93

Analysis of the components in our framework. Compared with other methods, the core components of our proposed method lie in the reflection-aware guidance network(RDNet) based on MaxRF, and the usage of a real-world dataset collected with the proposed pipeline as additional training data. We conduct experiments to evaluate the impact of essential components of our framework. We compare three models with different configurations: (i) RDNet + RRNet: the extra data from RRW is not involved in our training dataset, which aims to assess the effect of the extra data in our framework. (ii) RRNet + Extra Data: to assess the effect of the reflection-location guidance, only using reflection images as inputs. (iii) RDNet + RRNet + Extra Data: core components of our framework both are included. This is also our default model configuration.

The quantitative results are reported in Table 3. Augmented by the incorporation of supplementary data RRW, our approach exhibits a notable enhancement in performance across various reflection scenarios. This performance improvement is especially noticeable with the *Real* and *Nature* datasets. In contrast to the many controlled reflection scenes in SIR^2, these two datasets primarily encompass a range of wild reflection scenes. Therefore, powered by these two core innovations, our method can achieve the best de-reflection effects and exhibit superior generalization across real-world reflection scenarios.

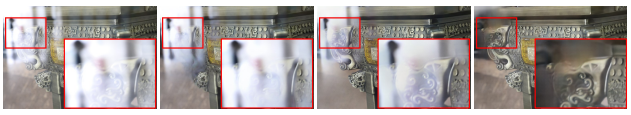

Figure 8. Failure case in the saturated reflection scene.

Extension of our proposed methods. The proposed innovations also could be extended to the previous methods. In the ablation study, we apply the reflection location guidance and the additional data expansion for methods [46, 53]. For the former, we cascade the RDNet before the network framework of the previous method. This is in line with our own framework, thereby constructing a two-stage network structure with reflection-aware guidance. As for the latter, we further supplement the cascaded framework by incorporating RRW data as an additional data expansion. The experimental results can be found in Table 4. This indicates our proposed innovations benefit pioneer methods as well, enhancing the generalization in the real reflection scenarios.

6. Conlusion

In this paper, we revisit the physical formulation of the reflection degradation imaging. This motivates us to propose a more applicable pipeline for real-world reflection data collection. Our pipeline is conducted in the video collection mode, enabling us to collect a large-scale, high-quality reflection training dataset at a lower cost, named RRW. Furthermore, we propose MaxRF, which explicitly identifies the locations of reflection distortions from aligned reflection image pairs, and develop a location-aware guidance framework for SIRR. With the support of our customized framework and RRW, our solution achieves superior performance against SOTA methods in real-world benchmarks. Meanwhile, we also suggest that these innovations could enhance the reflection removal performance of previous methods, hoping to inspire future research in this field.

Limitations. Our method may encounter challenges in scenarios with saturated reflections. The intensity of these saturated reflections is extremely high, making the underlying transmission contents nearly invisible in these saturated regions. Figure 8 illustrates this limitation. Although our method can mitigate these reflections to a certain degree, residual artifacts may remain. Given that the transmission contents loss occurs in the saturated regions, which often constitute a portion of the image, we intend to integrate semantic information to further enhance the restoration performance in future work.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207 and 62276243.

References

[1] Amgad Ahmed, Suhong Kim, Mohamed Elgharib, and Mohamed Hefeeda. User-assisted video reflection removal. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pages 122–131, 2021. 2

[2] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Susstrunk. Single image reflection suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4498–4506, 2017. 3

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022. 4

[4] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5017–5026, 2021. 2, 3, 4, 5, 7, 8

[5] 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*, pages 3238–3247, 2017. 3

[6] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. 4

[7] 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*, pages 2187–2194, 2014. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[9] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 5, 7

[10] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13138–13147, 2023. 3, 5, 7

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 7

[12] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5164–5173, 2020. 3, 7

[13] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[14] Chenyang Lei and Qifeng Chen. Robust reflection removal with reflection-free flash-only cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14811–14820, 2021. 3

[15] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1750–1758, 2020. 1, 2, 3, 4, 5

[16] Chenyang Lei, Xuhua Huang, Chenyang Qi, Yankun Zhao, Wenxiu Sun, Qiong Yan, and Qifeng Chen. A categorized reflection removal dataset with diverse real-world scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3040–3048, 2022. 1, 3

[17] Chenyang Lei, Xudong Jiang, and Qifeng Chen. Robust reflection removal with flash-only cues in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[18] 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*, 29(9):1647–1654, 2007. 3

[19] 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*, pages 3565–3574, 2020. 1, 2, 3, 4, 5, 7

[20] Tingtian Li, Yuk-Hee Chan, and Daniel PK Lun. Improved multiple-image-based reflection removal algorithm using deep neural networks. *IEEE Transactions on Image Processing*, 30:68–79, 2020. 2

[21] 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*, pages 2752–2759, 2014. 3

[22] Yu Li, Ming Liu, Yaling Yi, Qince Li, Dongwei Ren, and Wangmeng Zuo. Two-stage single image reflection removal with reflection-aware guidance. *Applied Intelligence*, pages 1–16, 2023. 3

[23] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Semantic guided single image reflection removal. *arXiv preprint arXiv:1907.11912*, 2019. 3

[24] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 182–199. Springer, 2020. 1

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[26] Daiqian Ma, Renjie Wan, Boxin Shi, Alex C Kot, and Ling-Yu Duan. Learning to jointly generate and separate reflections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2444–2452, 2019. 3

[27] Shree K Nayar, Xi-Sheng Fang, and Terrance Boult. Separation of reflection components using color and polarization. *International Journal of Computer Vision*, 21(3):163–186, 1997. 2

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

Learned dual-view reflection removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3713–3722, 2021. 2

[29] BH Prasad, Lokesh R Boregowda, Kaushik Mitra, Sanjoy Chowdhury, et al. V-desirr: Very fast deep embedded single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2390–2399, 2021. 3

[30] Jiaxiong Qiu, Peng-Tao Jiang, Yifan Zhu, Ze-Xin Yin, Ming-Ming Cheng, and Bo Ren. Looking through the glass: Neural surface reconstruction against high specular reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20823–20833, 2023. 1

[31] Hamed RahmaniKhezri, Suhong Kim, and Mohamed Hefeeda. Unsupervised single-image reflection removal. *IEEE Transactions on Multimedia*, 2022. 3

[32] Green Rosh, BH Pawan Prasad, Lokesh R Boregowda, and Kaushik Mitra. Deep unsupervised reflection removal using diffusion models. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2045–2049. IEEE, 2023. 3

[33] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 5

[34] Yoav Y Schechner, Joseph Shamir, and Nahum Kiryati. Polarization and statistical analysis of scenes containing a semireflector. *JOSA A*, 17(2):276–284, 2000. 2

[35] 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*, pages 3193–3201, 2015. 3

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[37] Sudipta N Sinha, Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski. Image-based rendering for scenes with reflections. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 1

[38] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4

[39] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017. 1, 2, 3

[40] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Wen Gao, and Alex C Kot. Region-aware reflection removal with unified content and gradient priors. *IEEE Transactions on Image Processing*, 27(6):2927–2941, 2018. 2, 4

[41] 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*, pages 4777–4785, 2018. 3

[42] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Corrn: Cooperative reflection removal network. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):2969–2982, 2019.

[43] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, and Alex C Kot. Reflection scene separation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2398–2406, 2020. 3

[44] Mengyi Wang, Xinxin Zhang, Yongshun Gong, and Yilong Yin. Personalized single image reflection removal network through adaptive cascade refinement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8204–8213, 2023. 7

[45] 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

[46] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019. 1, 3, 4, 7, 8

[47] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019. 3, 7

[48] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG)*, 34(4): 1–11, 2015. 3

[49] Tao Yan, Helong Li, Jiahui Gao, Zhengtian Wu, and Rynson WH Lau. Single image reflection removal from glass surfaces via multi-scale reflection detection. *IEEE Transactions on Consumer Electronics*, 2023. 4

[50] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the european conference on computer vision (ECCV)*, pages 654–669, 2018. 3, 7

[51] Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, and Weiyu Xu. Fast single image reflection suppression via convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8141–8149, 2019. 7

[52] Takahiro Yano, Masao Shimizu, and Masatoshi Okutomi. Image restoration and disparity estimation from an uncalibrated multi-layered image. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 247–254. IEEE, 2010. 4

[53] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 1, 2, 3, 4, 5, 8

[54] Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, and Alex C Kot. Single image reflection removal with absorption effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13395–13404, 2021. 3

[55] Yurui Zhu, Xueyang Fu, Zheyu Zhang, Aiping Liu, Zhiwei Xiong, and Zheng-Jun Zha. Hue guidance network for single image reflection removal. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3