A Dual-Stream Diffusion Model with Physically-Based Rendering for Single Image Reflection Removal

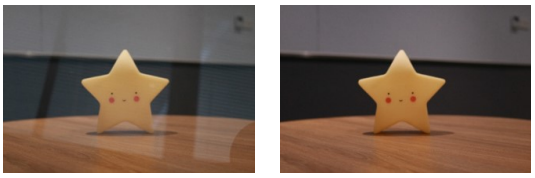
Cheng-Wei Hsu and Ming-Sui Lee
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
E-mail: r11922125@csie.ntu.edu.tw, mslee@csie.ntu.edu.tw

***Abstract*—Reflections are ubiquitous in our daily lives, making it inevitable for standard photographic equipment to capture unwanted reflected objects when taking images. Reflections degrade image aesthetics and can impair downstream vision tasks, making their removal from a single image a long-standing research focus. Although deep learning-based methods have made significant progress compared to traditional approaches in recent years, their performance is still limited by two fundamental issues: overly simplified reflection model assumptions and the domain gap between synthetic and real-world reflection images. A diffusion model-based approach is introduced to reduce dependency on assumptions, using a dual-stream network architecture to simultaneously predict residuals and reflection layers, thereby enhancing the diffusion model's ability to capture complex data distributions. Additionally, we employ physically based rendering techniques to generate the necessary training datasets, narrowing the gap between real-world images and synthetic data. Experimental results on the benchmark data demonstrate that the proposed model achieves performance comparable to state-of-the-art methods.**

I. Introduction

Reflections from windows, glass, and other reflective materials are unavoidable in real-world environments. When capturing photographs, especially in urban environments or indoor settings with large windows, reflections often appear unintentionally. This results in a composite image (Fig. 1a) where the scene of interest (transmission layer T), Fig. 1b) is mixed with unwanted reflections (reflection layer RRR), degrading image quality and clarity. Image segmentation, object detection, and face recognition systems may struggle to process such images accurately, as reflections obscure details, create false edges, or introduce artifacts that confuse AI models. This reduces accuracy in tasks important for applications like autonomous driving, surveillance, and image-based medical diagnostics. Removing reflections offers key benefits: it enhances visual quality, improves content visibility, and reveals previously obscured details. Technically, cleaner images improve computer vision performance across tasks. Additionally, reflection removal can help extract useful information hidden within reflections, offering unexpected value.

Given these challenges and potential benefits, Single Image Reflection Removal (SIRR) has emerged as an important area of research in computer vision. The primary goal of SIRR is to develop techniques capable of separating the transmission layer (reflection-free image) using only a single input image without any additional clues. This task is particularly challenging because it belongs to an ill-posed problem – there are theoretically infinite ways to decompose a mixed image into its constituent layers without additional information. SIRR research has explored various assumptions to represent the relationship between the captured image **I**, transmission layer **T**, and reflection layer **R**. The foundational model [1] is a simple linear combination while a more generalized form incorporates alpha-matting maps [2], to distinguish the weight for each pixel. However, these simple linear combination models may oversimplify the problem in the real world. Therefore, [3] has introduced a residual term to model the intricate effects across diverse scenarios. More complex assumptions may improve performance in diverse real-world conditions, but they also contribute to increased complexity in problem formulation and the model design. Moreover, the current mainstream deep learning-based methods for SIRR inevitably leverage synthetic data to facilitate training, due to the scarcity of real-world training pairs. Synthetic datasets are generally constructed using the aforementioned approximate assumptions. However, this leads to a pronounced domain gap between real-world images containing reflections and the synthetic data used for training.



(a) Reflection image I (b) Transmission layer T

Fig. 1: An example of reflection and reflection-free image pair.

Motivated by the effectiveness of diffusion-based image restoration models, Residual Denoising Diffusion Models (RDDM) [4], an extension of diffusion models, is applied to the SIRR task. The proposed approach directly predicts the residual between the reflected and reflection-free images, reducing reliance on SIRR assumptions. An additional branch is

979-8-3315-7206-8/25/$31.00

introduced to predict the reflection layer, which facilitates the RDDM in modeling the data distribution of the transmission layer. To mitigate the limitations of synthetic training data, we adopt physically based rendering techniques [5], which effectively narrow the domain gap between real-world and synthetic data, thus improving the model's generalization capability. In sum, this paper makes the following contributions:

- A diffusion model-based approach that reduces reliance on assumptions, demonstrating its feasibility and effectiveness for SIRR tasks.
- By incorporating a dual-stream network modification and physically based rendering training data, the proposed method significantly improves the diffusion model's performance in learning the transmission layer's data distribution, achieving competitive performance to state-of-the-art models.

II. Related Work

This section reviews prior work on Single Image Reflection Removal (SIRR), covering deep learning-based methods and their advancements and limitations. It also discusses common model assumptions and their impact on the effectiveness and accuracy of reflection removal. Finally, the emerging role of diffusion models in image restoration is examined, highlighting their potential for improving image quality and removing artifacts.

A. Deep Learning Methods on SIRR

Image reflection removal has been extensively studied, with methods broadly classified into multi-image and single-image approaches. This work focuses on the latter. Early methods often relied on gradient-based constraints [6]–[8], sparse priors [7], [9], and sometimes manual annotations [1] to mitigate the ill-posed issue on SIRR. While effective in specific scenarios, these approaches were limited by their reliance on specific assumptions and handcrafted features, which could be violated in complex real-world situations. As deep learning has developed, these approaches [3], [10]–[14] have begun to surpass the performance of traditional methods, dominating the SIRR task. CEILNet [10] introduced a two-stage network approach, first estimating an edge map and then using it to guide the reconstruction of the transmission layer. This method imposed a relative smoothness prior on reflection layers and combined them additively with transmission layers. However, it struggled to capture high-level semantic information, which limited its performance in complex scenarios. Building on this foundation, Zhang et al. [11] incorporated HyperColumn features [15] from a pre-trained network along with perceptual loss to capture semantic knowledge and also introduced an exclusivity loss to decrease the overlapping edge between T and R. ERRNet [12] further advanced the field by using multi-scale channel-wise context and also leveraged misaligned real-world image pairs to enhance performance. However, it did not explicitly predict the reflection layer, potentially overlooking valuable information. Recognizing the importance of both transmission and reflection components, subsequent approaches like DMGN [13] and YTMT [14] adopted dual-stream frameworks. The former utilizes an attention mechanism to generate a gating mask, which controls information flow; the latter implements negative ReLU [16] to leverage mutual information between the reflection and transmission layers.

B. Assumptions of Reflection Model

Several research has explored various assumptions to represent the relationship between $\mathbf{I}$, $\mathbf{T}$, and $\mathbf{R}$. This additive model $\mathbf{I} = \mathbf{T} + \mathbf{R}$ [1], while popular for its simplicity, has been expanded upon to address real-world complexities. Some researchers [17], [18] introduced scalar coefficients $\mathbf{I} = \alpha\mathbf{T} + \beta\mathbf{R}$ to account for potential weakening of layers due to diffusion and other factors. To handle phenomena like overexposure that violate the linear model, more sophisticated approaches incorporated alpha-matting maps $\mathbf{I} = \mathbf{W} \circ \mathbf{T} + (\mathbf{1} - \mathbf{W}) \circ \mathbf{R}$ [2]. Building on these concepts, DSRNet [3] has proposed a more generalized form of the superimposition process which introduces a residual term, resulting in the equation $\mathbf{I} = \mathbf{T} + \mathbf{R} + \Phi(\mathbf{T}, \mathbf{R})$, where $\Phi(\cdot, \cdot)$ denotes a collection of functions that can model residual effects across diverse scenarios via a learnable residue module (LRM) which used to estimate the remaining residual term from the fused feature of the dual stream. However, these assumptions still fall short of accurately representing real-world phenomena. Simplified assumptions in network design can lead to error accumulation and degrade image quality. Synthetic data from such assumptions [3], [10], [14] often reduces model performance on real-world images. To address this, Kim et al. [5] proposed a physically based rendering method that simulates real-world glass and lens effects using physical engines. This approach models complex interactions between transmission and reflection layers, incorporating accurate light behavior, depth-dependent blur, and refraction factors often overlooked in conventional synthesis methods.

C. Diffusion Models for Image Restoration

Diffusion models [19] are a category of generative models designed to convert noise into structured data through iterative denoising steps. They operate by incrementally introducing noise to data and subsequently learning to reverse this process, enabling the generation of new samples. These models have garnered considerable interest for their capacity to generate high-quality outputs and maintain a stable training process. In the field of image restoration, diffusion models have gradually become a promising approach, particularly excelling in low-light vision tasks such as super-resolution, image inpainting, image restoration, and shadow removal. Saharia et al. [20] demonstrated the superiority of denoising diffusion probabilistic models over GAN-based methods in image super-resolution, while RePaint [21] introduced an innovative mask-agnostic approach for free-form inpainting, yielding photo-realistic results. LFG-Diffusion [22] took a novel approach by incorporating a latent feature space that captures perceptual shadow-free priors to guide the diffusion

model. Unlike previous diffusion model-based image restoration methods that use conditional generation to restore images, Residual Denoising Diffusion Models (RDDM) [4] propose incorporating the residual (difference between the degraded image and the clean image) into the diffusion process. RDDM offers a more interpretable way to establish the relationship between degraded and clear images and has been applied to various restoration tasks such as shadow removal, low-light enhancement, deblurring, and deraining.

III. Method

In this section, a diffusion model-based method for SIRR using Residual Denoising Diffusion Models is introduced. A dual-stream network is proposed for predicting the reflection layer, including both the loss function and network architecture. Finally, we present our method for constructing a PBR dataset specifically designed for SIRR tasks.

Based on the derivation of RDDM [4], we define $x_{res} = x_{in} - x_0$ is the difference between reflection image x_{in} and reflection-free image x_0 and finally the training objective of RDDM [4] is obtained as follow:

$$\mathcal{L}_{res} := \mathbb{E}_{t\sim U[0,T]}[\|x_{res} - x^{\theta}_{res}(x_t, t, x_{in})\|_1] \tag{1}$$

A. Additional branch of reflection

As derived in the discussion of RDDM [4] above, the entire process only considers the prediction of the residual, which corresponds to the data distribution of transmission layer (reflection-free image). However, for the task of reflection removal, predicting the reflection layer also plays a critical role [3], [14], [23]. Intuitively, there may be a high correlation between the content of the residual and the reflection, as both contain meaningful information for each other. Therefore, utilizing information about the reflection as guidance can not only help the model predict more precise residuals but also increase the interpretability of the model. The loss of reflection layer is defined as follows:

$$\mathcal{L}_R := \mathbb{E}_{t\sim U[0,T]}[\|x_R - x^{\theta}_R(x_t, t, x_{in})\|_1] \tag{2}$$

The network architecture is based on the implementation of [24] which employs a multi-layer Unet [25] design, incorporating residual and attention blocks. The model's input comprises a 6-dimensional feature obtained by concatenating x_t and x_{in} along the channel dimension, along with the time condition t. To simultaneously predict residuals and reflections, we added two sets of independently weighted convolution and residual blocks to the front and back of the Unet, transforming the architecture into a dual-stream network like Fig. 2. All other parameters within the Unet are shared, which reduce the memory and computational burden of the dual-stream network to some extent.

B. Loss Function

In addition to the reconstruction loss of the residual and reflection layers, we also incorporate a gradient loss into our model to preserve edge information and fine details in the both transmission and reflection layer. The overall loss is defined as:

$$\begin{aligned}\mathcal{L}_{all} = \gamma_1\mathcal{L}_{res} + \gamma_2\mathcal{L}_R + \\ \gamma_3\mathbb{E}_{t\sim U[0,T]}[||\nabla x_{res} - \nabla x^{\theta}_{res}||_1] + \\ \gamma_4\mathbb{E}_{t\sim U[0,T]}[||\nabla x_R - \nabla x^{\theta}_R||_1]\end{aligned} \tag{3}$$

where $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.75, 0.25, 0.25, 0.25)$. Finally, the model parameters are updated using Eq. (3).

C. Physically Based Rendering Dataset Generation

The proposed method require a sizable training data to simulate the data distribution of reflection-free images. However, paired images of reflection and reflection-free are very rare, necessitating the use of synthetic methods to augment the data. Overly simplified synthetic formulas can result in significant discrepancies between the synthetic results and real data, which greatly affects models like diffusion models that simulate the original data distribution. Therefore, we ultimately adopted a physically based rendering synthesis method [5].

The process of generating physically based rendering (PBR) synthetic data is described as follows: we randomly select two images from the DIODE dataset [26] and position them at ±1 meters to serve as the foreground and background. When rendering the reflection image, the camera faces and focuses on the foreground, with the 10 millimeters thick glass (index of refraction = 1.5) placed at +20 centimeters like Fig. 3. The transmission and reflection layer are then rendered by respectively removing the light sources of the background and foreground. Examplar rendering results shown in Fig. 4.

IV. Experimental Results

A. Implementation Details

Model training was conducted with a batch size of 5 for around 1.4k epochs, and input were cropped to 384×384 pixels. The optimization process utilized the Adam optimizer [27] with a learning rate of 8×10^{-5} and Exponential Moving Average (EMA) algorithm for parameters update. During inference, the backward diffusion steps are set to 2. All experiments were conducted on single NVIDIA RTX A6000 GPU and the implementation leveraged the PyTorch framework.

B. Dataset and Evaluation Metrics

We follow the experiment setting in previous work [3], [23] benchmark in three real-world dataset. (1) Nature dataset [23] includes 200 training and 20 testing pairs (reflection image and reflection-free image). (2) Real dataset [11] includes 89 training and 20 testing pairs. (3) SIR wild [28] includes 55 testing pairs. Our synthesized training dataset comprises 800 images created from the DIODE dataset [26] using the physically based rendering technique described in our methodology III-C. The evaluation metrics contain PSNR and SSIM, both computed in the RGB color space.

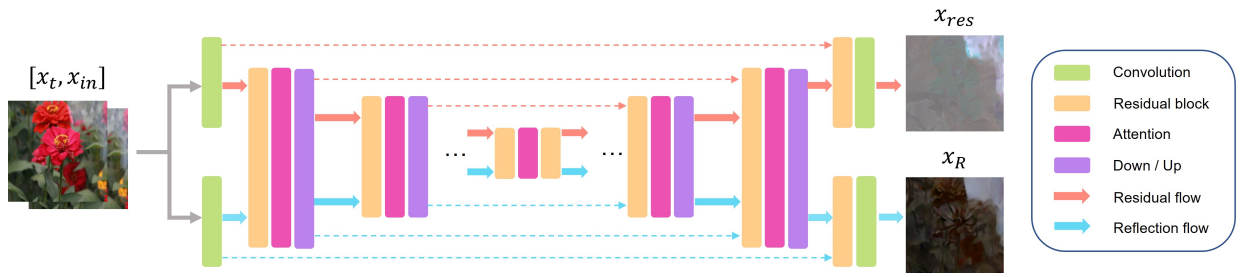


Fig. 2: The diagram is an overview of network structure.

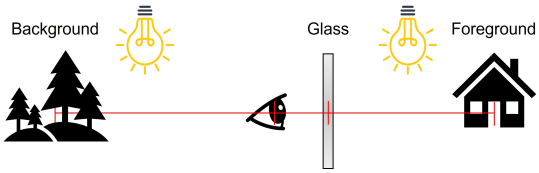


Fig. 3: An overview of the rendering setting.

Fig. 4: Examples of PBR synthetic dataset.

C. Comparison with State-of-the-art Methods

Quantitative Comparison. We compare our method with state-of-the art methods, involving IBCLN [23], DMGN [13], YTMT [14] and DSRNet [3] as shown in Table I. It turns out that our method demonstrates superior performance across diverse reflection removal scenarios, particularly excelling in the Nature dataset due to the availability of sufficient training data. It shown that our dual stream diffusion model effectively captures the data distribution of real world reflection-free images. Notably, our method achieves competitive results on the Real20 and SIR Wild datasets compared to previous methods [3], [14] those using pretrained VGG [29] features that already containing plentiful knowledge. This highlights the model's robustness and generalization ability, attributed to the high-quality PBR synthetic data.

Visual Comparison. Fig. 5 clearly demonstrates the superior performance of our method across diverse reflection removal scenarios. Our approach consistently maintains the clarity and color fidelity of the transmission layer while effectively removing reflections, a balance that other methods struggle to achieve. Notably, some competing methods, while removing reflections, introduce significant color distortion. For example, in the first and second rows of Fig. 5, DSRNet [3] and IBCLN [23], while successfully removing most of the reflections, have inadvertently introduced some color changes that negatively impact the visual experience. The color distortion issue is likely due to these models being trained on non physically based rendering (non PBR) data, as we demonstrated in our ablation study. Across all scenarios, our approach consistently produces results that closely align with the ground truth, effectively removing reflections without compromising the original scene's fidelity.

TABLE I: Quantitative comparison with other SOTA model on three real-world benchmark datasets.

Method	Testing Data					
	Nature		Real20		SIR Wild	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
IBCLN [23]	<u>23.91</u>	0.784	21.70	0.762	24.70	0.885
DMGN [13]	20.83	0.764	20.71	0.763	23.81	0.835
YTMT [14]	20.93	0.773	22.94	**0.803**	<u>25.23</u>	<u>0.890</u>
DSRNet [3]	21.51	<u>0.787</u>	**23.51**	0.794	24.64	**0.893**
Ours	**24.97**	**0.811**	<u>23.17</u>	<u>0.796</u>	**25.29**	<u>0.890</u>

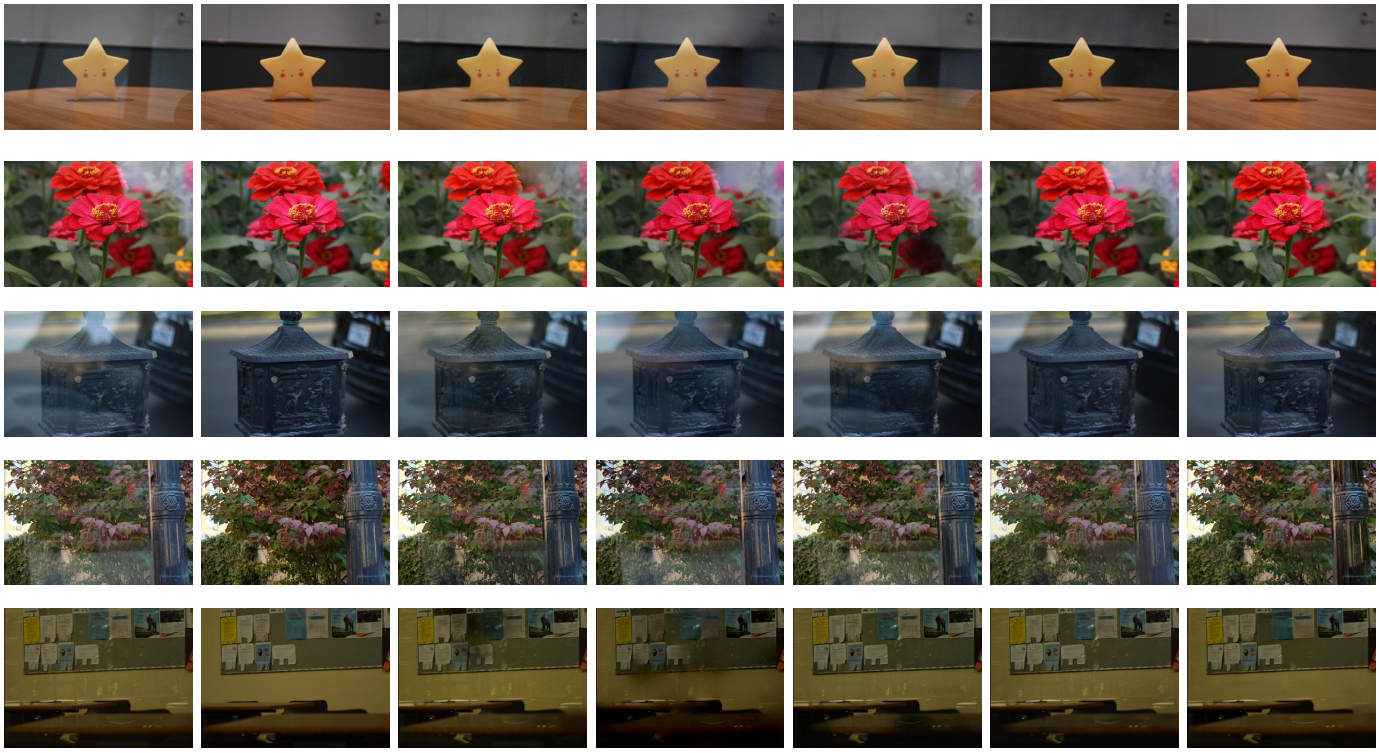
D. Ablation Study

The study focuses on two key aspects: the impact of using physically based rendering (PBR) for synthetic training data generation, and the effect of incorporating explicit reflection prediction in the model architecture. All quantitative results are shown in Table II.

TABLE II: Ablation study on different configurations.

Configuration			Testing Data					
Non PBR Synth.	PBR Synth.	Predict Reflection	Nature		Real20		SIR Wild	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
			<u>24.45</u>	0.801	20.79	0.766	24.88	0.884
✓			23.00	0.798	20.24	0.734	24.03	0.882
	✓		24.06	**0.814**	21.61	<u>0.780</u>	<u>25.03</u>	<u>0.888</u>
✓		✓	24.32	0.809	<u>21.92</u>	0.766	24.61	0.883
	✓	✓	**24.97**	<u>0.811</u>	**23.17**	**0.796**	**25.29**	**0.890**

Physically Based Rendering Dataset. For the non PBR approach, we followed the synthetic formulas used in previous works [3]. While incorporating non PBR synthetic data unexpectedly diminishes performance across all datasets, PBR



(a) Input (b) GT (c) IBCLN [23] (d) DMGN [13] (e) YTMT [14] (f) DSRNet [3] (g) Ours

Fig. 5: Visual comparisons between input reflection image, ground truth and different approaches.

synthetic data, in contrast, leads to substantial improvements in results. This shows that the choice between PBR and non PBR approaches has a significant impact on diffusion model-based methods.

Reflection Branch. Whether using PBR or non PBR synthetic data, the addition of reflection prediction consistently enhances performance across all datasets, as evidenced by comparing the models with and without this feature. This suggests that explicitly modeling the reflection component allows the network to more effectively separate it from the transmission layer, resulting in improved overall reflection removal quality.

Overall Comparison. Table II clearly demonstrates that the combination of PBR synthetic data and reflection prediction yields the best performance across all datasets, while Fig. 6 showcases a visual comparison of different model configurations across several scenes. Although the baseline model shows some ability to remove reflections, it often leaves artifacts or residual reflections (indicated by green bounding boxes). In contrast, the non PBR configuration performs poorly, frequently introducing new artifacts or distortions (marked by red bounding boxes). The PBR configuration, however, shows noticeable improvements with fewer artifacts and better reflection removal, though some challenging areas remain. Interestingly, adding reflection prediction to the non PBR configuration yields mixed results. Our proposed method consistently demonstrates the best performance across all scenes, effectively removing reflections while preserving original image details and minimizing artifacts. This visual comparison strongly supports the quantitative results from the ablation study table, illustrating the cumulative benefits of using PBR synthetic data and incorporating reflection prediction in the model architecture, thus underscoring the importance of both data quality and model design in addressing the challenge of reflection removal.

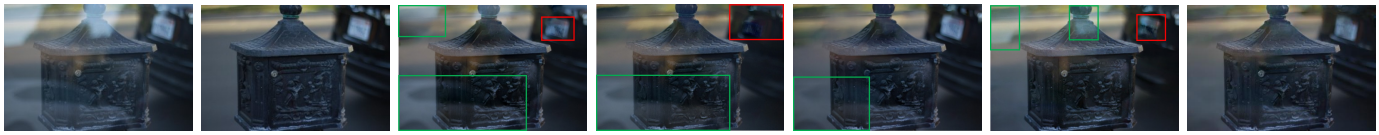
V. Conclusion

This paper introduces a diffusion model-based approach for Single Image Reflection Removal (SIRR), addressing key challenges through a dual-stream network and physically based rendering for synthetic data. By reducing reliance on SIRR-specific assumptions, the proposed method enables more flexible and accurate reflection removal. Experiments on real-world datasets demonstrate strong generalization and competitive performance, even with limited training data. This work highlights the potential of diffusion models in SIRR and paves the way for future applications in broader image restoration tasks.

References

[1] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *PAMI*, vol. 29, no. 9, pp. 1647–1654, 2007.

[2] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *CVPR*, 2019, pp. 3771–3779.



(a) Input (b) GT (c) Baseline (d) Non-PBR (e) PBR (f) Non-PBR + Predict R (g) Ours

Fig. 6: Visual comparisons of ablation study. Green bounding boxes denote the remaining reflection. Red bounding boxes denote the critical artifacts or distortions.

[3] Q. Hu and X. Guo, "Single image reflection separation via component synergy," in *ICCV*, 2023, pp. 13 138–13 147.

[4] J. Liu, Q. Wang, H. Fan, Y. Wang, Y. Tang, and L. Qu, "Residual denoising diffusion models," in *CVPR*, Jun. 2024, pp. 2773–2783.

[5] S. Kim, Y. Huo, and S.-E. Yoon, "Single image reflection removal with physically-based training images," in *CVPR*, 2020, pp. 5164–5173.

[6] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *CVPR*, 2014, pp. 2752–2759.

[7] A. Levin, A. Zomet, and Y. Weiss, "Separating reflections from a single image using local features," in *CVPR*, IEEE, vol. 1, 2004, pp. I–I.

[8] Y. Li and M. S. Brown, "Exploiting reflection change for automatic reflection removal," in *ICCV*, 2013, pp. 2432–2439.

[9] N. Arvanitopoulos, R. Achanta, and S. Susstrunk, "Single image reflection suppression," in *CVPR*, 2017, pp. 4498–4506.

[10] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *ICCV*, 2017, pp. 3238–3247.

[11] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *CVPR*, 2018, pp. 4786–4794.

[12] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *CVPR*, 2019, pp. 8178–8187.

[13] X. Feng, W. Pei, Z. Jia, F. Chen, D. Zhang, and G. Lu, "Deep-masking generative network: A unified framework for background restoration from superimposed images," *TIP*, vol. 30, pp. 4867–4882, 2021.

[14] Q. Hu and X. Guo, "Trash or treasure? an interactive dual-stream strategy for single image reflection separation," *NeurIPS*, vol. 34, pp. 24 683–24 694, 2021.

[15] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015, pp. 447–456.

[16] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[17] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Crrn: Multi-scale guided concurrent reflection removal network," in *CVPR*, 2018, pp. 4777–4785.

[18] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *ECCV*, 2018, pp. 654–669.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[20] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *PAMI*, vol. 45, no. 4, pp. 4713–4726, 2022.

[21] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *CVPR*, 2022, pp. 11 461–11 471.

[22] K. Mei, L. Figueroa, Z. Lin, Z. Ding, S. Cohen, and V. M. Patel, "Latent feature-guided diffusion models for shadow removal," in *WACV*, 2024, pp. 4313–4322.

[23] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," *arXiv preprint arXiv:1911.06634*, 2019.

[24] lucidrains, Adversarian, AlejandroSantorum, and nilsleh, *Denoising-diffusion-pytorch*, http://github.com/lucidrains/denoising-diffusion-pytorch, 2022.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[26] I. Vasiljevic, N. Kolkin, S. Zhang, *et al.*, "DIODE: A Dense Indoor and Outdoor DEpth Dataset," *CoRR*, vol. abs/1908.00463, 2019. [Online]. Available: http://arxiv.org/abs/1908.00463.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *ICCV*, 2017.

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