

Diffusion-based Dual-view Reflection Removal

Tianyi Xu¹, Zifeng Wang¹, Boyang Lv¹, Shuchen Weng^{2,1}, and Boxin Shi^{1†}

¹Peking University ²Beijing Academy of Artificial Intelligence
photon@stu.pku.edu.cn shiboxin@pku.edu.cn

Abstract. Reflection removal aims to recover clean images from photographs degraded by glass, a crucial problem in mobile photography. Single-image methods suffer from the fundamental ill-posedness of de-composition and often rely on specialized hardware to enhance robustness. We propose a user-friendly dual-view diffusion framework that effectively leverages complementary viewpoint information for ubiquitous dual-camera devices. Our approach introduces frame-aware spatial attention to capture implicit cross-view correlations and integrates camera pose as conditional embeddings for geometric guidance. We contribute DUALREF, a dataset for reflection removal comprising semi-synthetic pairs and real-world dual-camera sequences with ground truth. Experiments demonstrate over 10% PSNR improvement on real-world data, significantly outperforming state-of-the-art methods.

Keywords: Reflection removal · Diffusion model · Dual-view

1 Introduction

Reflection removal aims to recover clean transmission layers from images degraded by unwanted reflections through glass surfaces [64, 65]. This problem is ubiquitous in everyday photography scenarios [5, 52, 63]—from museum exhibits and vehicle windows to architectural facades—significantly impacting image quality in mobile photography applications. Among various capture setups, single-image reflection removal (SIRR) is the most user-friendly, requiring minimal operational constraints. However, SIRR is inherently ill-posed: decomposing a single mixture image $\mathbf{M}$ into transmission $\mathbf{T}$ and reflection $\mathbf{R}$ [64, 65] components admits infinite valid solutions, highlighting the difficulty of achieving robust reflection removal from a single image [11, 17, 27, 29, 30, 33, 39, 87, 90, 92]. As shown in the view 1 input image in Figure 1, even human observers may find it challenging to distinguish between reflection and transmission, especially with structurally complex or high-intensity reflections.

To mitigate this ill-posedness, researchers have explored various auxiliary constraints with varying levels of operational complexity. Hardware-based methods utilize polarization filters [44], near-infrared imaging [22, 23], flash/no-flash

[†] Corresponding author. ¹ State Key Laboratory of Multimedia Information Processing, National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

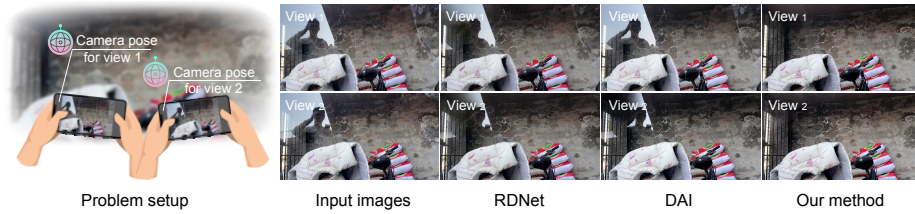


Fig. 1: Dual-view reflection removal with camera pose guidance. **Left:** Problem setup showing sequential capture of two viewpoints with camera poses tracked via device gyroscope. **Right:** Comparison of reflection removal results on challenging real-world data with overexposed reflections. The upper and bottom rows display views 1 and 2, respectively. While state-of-the-art single-image methods, such as RDNet [90] and DAI [27], fail to remove intense reflections when processing each view independently, our dual-view diffusion approach jointly processes both views to achieve effective reflection removal while preserving background fidelity.

pairs [21, 38], or panoramic capture [24, 25] to provide reliable physical cues, albeit at the cost of specialized equipment and complex user operations, which limits their practical deployment. Recent approaches incorporate language guidance [26, 91] to offer greater flexibility while still requiring user input to describe desired outputs. Another branch of research lies in multi-view reflection removal, which leverages image sequence input [1, 46, 58] based on deliberate camera movements, yet requiring capturing and processing multiple frames, significantly increasing both computational complexity and user effort. Modern smartphones equipped with dual-camera systems and Live Photo capabilities [2] can capture image pairs seamlessly, often without the user’s awareness. Hence, we simplify the multi-frame setting to a user-friendly dual-view reflection removal.

However, transitioning to a dual-view setup drastically reduces the redundancy in observations across multi-frame sequences. To compensate, a model must fully exploit cross-view correspondences. Previous multi-view methods typically build upon explicit optical flow estimation [46] or 3D reconstruction [86]. Yet, these explicit techniques are highly susceptible to severe reflection interference and frequently fail in a constrained dual-view setting. Furthermore, because the physical transmission and virtual reflection layers reside at different depths, both exhibit distinct depth-induced parallax shifts. A robust framework must learn to disentangle these overlapping motion patterns while adapting to the varying viewpoint baselines inherent to mobile capture systems.

To address these challenges, we present an end-to-end dual-view diffusion framework that effectively separates reflections from transmissions without relying on fragile explicit correspondence matching or rigid 3D constraints. First, we propose a frame-aware spatial attention mechanism that implicitly captures cross-view correlations. Unlike naive joint attention, which indiscriminately mixes features and amplifies mutual noise across views, our design enforces a directed attention pattern: noisy transmission queries attend exclusively to clean mixture conditions. This elegantly bypasses error-prone optical flow and strictly prevents

cross-view noise leakage, all while seamlessly reusing powerful pre-trained generative priors. Second, inspired by how human observers shift their viewpoint to distinguish physical depth from reflections, we introduce a camera pose encoding. Rather than enforcing computationally expensive and brittle 3D hard constraints, we inject the relative camera pose as a soft conditional embedding. This lightweight global prior guides the network to disentangle depth-induced parallax motion, while exhibiting strong robustness against the inherent drift and noise of ubiquitous smartphone IMU/gyroscope sensors, as shown in Figure 1.

To enable training and evaluation of our approach, we construct DUALREF, a comprehensive dataset addressing the scarcity of dual-view reflection removal data. DUALREF comprises semi-synthetic image pairs for training and testing, along with real-captured dual-view videos with reflections, which feature ground truth captured with a dual-camera dolly system and synchronized pose tracking, providing both training data and real-world evaluation benchmarks.

Our principal contributions are:

- A diffusion-based reflection removal method designed specifically for the dual-view setting by proposing:
 - a frame-aware attention mechanism for effective implicit cross-view feature correlation that prevents cross-view noise leakage and
 - a robust camera pose integration scheme through learned embeddings that disentangles depth-induced parallax motion.
- DUALREF dataset comprising semi-synthetic image pairs and real-world evaluation sequences with ground truth.

Extensive experiments demonstrate that our dual-view approach achieves remarkable results, surpassing single-image methods by over 10% in PSNR and comparable to complex multi-frame sequence methods.

2 Related work

Reflection removal. Deep learning approaches have substantially advanced SIRR [11, 12, 17, 27–30, 33, 39, 60, 64, 65, 73, 74, 82, 87, 89, 90]. Recently, diffusion-based methods [13, 26, 27, 33, 70, 87] that leverage powerful generative priors to handle ambiguous decomposition problems have emerged, with DAI [27] achieving single-step generation and L-DIFFER [26] incorporating language guidance. However, because SIRR is fundamentally ill-posed, these methods still struggle with complex real-world reflections. Researchers also leveraged auxiliary information to address this ambiguity at the cost of requiring specialized hardware: polarization [34, 37, 43, 44], NIR imaging [22, 23], flash/no-flash pairs [21, 35, 38, 71], dual-pixel cues [51, 85, 86], and wide field-of-view capture [19, 24, 25, 48].

Multi-view strategies exploit geometric cues from viewpoint variation. Early methods [16, 18, 81] demonstrated layer separation through motion analysis. Burst photography [9, 50] demands less capture effort but provides limited parallax. Image sequence methods using deliberate camera motion [1, 46, 58] increase parallax but introduce computational complexity during inference. Dual-view



Fig. 2: Overview of the DUALREF dataset. **Left:** Examples from DUALREF-R showing semi-synthetic dual-view pairs with realistic and often structurally complex or high-intensity reflections generated through video-based synthesis and physics-based rendering. **Center:** Representative examples from DUALREF-V demonstrating challenging real-world sequences with diverse scenes, strong reflection intensities, and complex reflection attributes. **Right:** Distribution of frame displacement between views in DUALREF-R, centered around 50 pixels but spanning from near-zero (burst-like) to large displacements, providing diverse parallax patterns for enhanced generalization.

setups offer a balance; however, existing methods still face challenges: optical flow accumulates alignment errors, while 3D reconstruction [52, 86] requires additional depth sensing. By contrast, our method implicitly builds cross-frame correspondence via the proposed frame-aware spatial attention mechanism.

Diffusion models for image restoration. Diffusion models [10, 20, 59] excel at learning complex data distributions through progressive denoising, demonstrating remarkable capabilities in addressing under-constrained inverse problems [31, 55]. Applications span diverse restoration tasks: motion deblurring [77], resolution enhancement [53, 68, 72], colorization [6, 7, 75, 76], content completion [42, 80], and quality improvement [69, 79]. Their ability to preserve fine details while resolving structural ambiguities makes them suited for reflection removal. Within a diffusion framework, we propose a novel cross-view attention mechanism that prevents cross-view noise leakage while effectively reusing powerful generative priors.

3 Dataset

Existing reflection removal datasets [4, 13, 39, 65, 83, 89] primarily focus on single-view images. Furthermore, current dual-view datasets [46] are not fully open-source, and multi-view datasets [9] lack sufficient scale, significantly limiting the development of multi-view approaches. To bridge this gap, we propose DUALREF, a comprehensive dual-view reflection removal dataset comprising: (1) **DUALREF-R**: 16,000 semi-synthetic pairs with diverse motion patterns and reflection intensities for training and evaluation; and (2) **DUALREF-V**: 195 real-world sequences captured using smartphones for evaluation. DUALREF compre-

hensively covers diverse reflection scenarios featuring various intensities, blurs, and geometric distortions.

3.1 Semi-synthetic data generation

We develop comprehensive synthesis methods that faithfully model real-world reflection physics. Our DUALREF-R dataset comprises 16,000 semi-synthetic pairs generated through two complementary approaches, with example images displayed in Figure 2 (left panel). The distribution of frame displacement (right panel) spans from near-zero to significant shifts, providing diverse parallax patterns for enhanced generalization.

Video-based synthesis: We leverage dense video sequences with camera trajectories, sampling frame pairs with controlled separation to ensure sufficient baseline while maintaining scene overlap. The synthesis process models realistic reflection-transmission mixing using the standard optical model with gamma correction, Gaussian blur, and ghosting masks. We collect 4,000 high-quality, pure-reflection images as reflection sources. The reflection layer exhibits view-dependent motion through homography warping with stochastic perturbation.

Physics-based rendering: We generate 4,000 dual-view pairs with precise ground truth through physically-based rendering. Our setup models realistic glass reflection scenarios (50% indoor-to-outdoor, 30% outdoor-to-indoor, 20% uniform environments) with realistic optical properties ($\text{IOR} = 1.43$) and 2-4cm baseline distances plus random perturbations to simulate handheld capture.

3.2 Real-world evaluation data

Evaluating real-world performance is critical, but is hindered by the lack of dual-view and video datasets with reflection ground truth, mainly because ensuring trajectory reproducibility is difficult. To address this, we introduce DUALREF-V, comprising 195 challenging real-world sequences (1200×650 resolution, 150 frames at 30fps, with synchronized 10Hz IMU & Gyroscope data). To obtain precise ground truth, we develop a trajectory-aligned capture system (Figure 3).

Our system has a programmable camera dolly equipped with a steerable wheel and high-precision IMU & Gyroscope, starting from a fixed docking station to ensure exact trajectory reproducibility. The collection protocol involves executing a programmed path to record the mixture sequence through a removable glass panel, and subsequently repeating the identical trajectory without the glass to capture the transmission-only ground truth. Finally, frame and pose data



Fig. 3: Trajectory-aligned video capture system for DUALREF-V. A programmable dolly ensures reproducible camera trajectories. It captures paired mixture and transmission-only ground truth by repeating paths with and without a removable glass panel, while an IMU & Gyroscope module records motion data.

adjustments are applied to align the pairs and compensate for minor trajectory deviations.

The center panel of Figure 2 showcases representative examples from DUAL-REF-V, demonstrating diverse scenes, strong reflection intensities, and complex reflection attributes. These challenging real-world sequences span diverse environments, including outdoor natural settings, vehicles, and architectural scenes with varying lighting conditions and reflection complexity. This setup enables the accurate quantitative evaluation of dual-view methods on real data with ground truth—a capability previously unavailable—and can be extended to support the evaluation of multi-view reflection removal methods, as well as Live Photo settings, through its continuous trajectory recording.

4 Method

As is shown in Figure 4, given a pair of mixture images $\mathbf{M}_1, \mathbf{M}_2$ and their corresponding camera poses $\mathbf{p}_1, \mathbf{p}_2$, our goal is to recover the clean transmission layers $\mathbf{T}_1, \mathbf{T}_2$ governed by the physical model $\mathbf{M}_i = \mathbf{T}_i + \mathbf{R}_i$. Building upon the Stable Diffusion v2.1 [53] architecture, we map the images into the latent space using a frozen VAE. Our modified U-Net then processes a unified joint representation, constructed by concatenating the noisy transmission latents and the clean mixture conditions from both views, and denoises both views simultaneously. We introduce our proposed frame-aware attention mechanism in Section 4.1 and camera pose integration scheme in Section 4.2.

4.1 Frame-aware spatial attention

Transitioning from multi-frame sequences to a dual-view setup inherently reduces the available redundancy in observations. To separate reflections from transmissions effectively, it is crucial to maximize the utility of cross-view motion correlations (i.e., parallax). Previous multi-view reflection removal methods [46, 81] typically rely on explicit optical flow to align frames. However, flow estimation is highly susceptible to errors when corrupted by severe reflection interference.

Within a diffusion framework, a natural alternative to explicit flow estimation is leveraging self-attention mechanisms to implicitly capture cross-view correspondences. Yet, directly adapting existing multi-view attention designs presents significant drawbacks. Naively extending standard self-attention to joint cross-view attention [14, 56] introduces an unrestricted bidirectional flow. In our dual-view setup, this inherently allows noisy latents to inappropriately attend to each other across views, mutually amplifying noise and disrupting the denoising process. Besides, adopting ResNet 3D layers [45] alters the inherent feature dimensions, which prevents the reuse of powerful pre-trained generative priors.

To overcome these limitations, we propose a frame-aware spatial attention mechanism that enables robust cross-view reasoning while strictly preventing noise propagation. As shown in Figure 4 (bottom right), our frame-aware spatial attention is applied to the 16×16 and 32×32 upsampling layers of the U-Net [54].

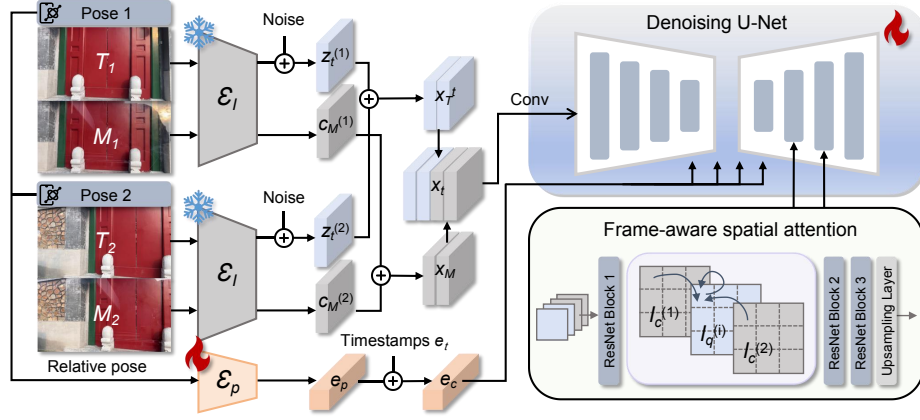


Fig. 4: Training pipeline of our dual-view diffusion framework. **Left:** Mixture and transmission images are encoded via frozen VAE. Transmission latents are noised while mixture latents remain clean, then concatenated channel-wise to form a 16-channel input to the modified U-Net [54]. Camera poses are converted to relative poses and encoded via an MLP, then combined with a timestep embedding for conditioning. **Bottom right:** Frame-aware spatial attention at 16×16 and 32×32 layers enables each noisy patch to attend to all mixture patches from both views, achieving cross-view exchange while reusing pretrained weights.

At these layers, the hidden features have been processed into 4 components corresponding to 2 views, each containing noisy and clean paths. Given the feature tensor $\mathbf{l} \in \mathbb{R}^{B \times 4C' \times H \times W}$ where C' is the feature dimension at this layer, we first separate it into query and condition components:

$$\mathbf{l}_q = \mathbf{l}[:, :2C', :, :], \quad \mathbf{l}_c = \mathbf{l}[:, 2C' :, :, :], \quad (1)$$

where $\mathbf{l}_q \in \mathbb{R}^{B \times 2C' \times H \times W}$ contains features from noisy transmission paths of both views, and $\mathbf{l}_c \in \mathbb{R}^{B \times 2C' \times H \times W}$ contains features from clean mixture paths. We then reshape these into a sequence format for attention computation:

$$\mathbf{l}_q \rightarrow \text{reshape} \rightarrow \mathbf{s}_q \in \mathbb{R}^{B \times 2HW \times C'}, \quad \mathbf{l}_c \rightarrow \text{reshape} \rightarrow \mathbf{s}_c \in \mathbb{R}^{B \times 2HW \times C'}. \quad (2)$$

We concatenate query and condition sequences along the sequence dimension and apply the pretrained projection matrices $\mathbf{P}_q, \mathbf{P}_k, \mathbf{P}_v$ from SD v2.1 [53]:

$$\mathbf{s}_{\text{concat}} = [\mathbf{s}_q; \mathbf{s}_c] \in \mathbb{R}^{B \times 4HW \times C'}, \quad (3)$$

$$[\mathbf{Q}_{\text{all}}, \mathbf{K}_{\text{all}}, \mathbf{V}_{\text{all}}] = [\mathbf{P}_q, \mathbf{P}_k, \mathbf{P}_v] \cdot \mathbf{s}_{\text{concat}}, \quad (4)$$

where each of $\mathbf{Q}_{\text{all}}, \mathbf{K}_{\text{all}}, \mathbf{V}_{\text{all}} \in \mathbb{R}^{B \times 4HW \times C'}$. Then we extract queries only from positions corresponding to noisy features, while keys and values come only from positions corresponding to clean conditions:

$$\mathbf{Q} = \mathbf{Q}_{\text{all}}[:, :2HW, :], \quad \mathbf{K} = \mathbf{K}_{\text{all}}[:, 2HW :, :], \quad \mathbf{V} = \mathbf{V}_{\text{all}}[:, 2HW :, :]. \quad (5)$$

This creates a directed attention pattern where each element in the query sequence (from noisy transmissions) attends to all elements in the key/value sequence (from clean mixtures):

$$\text{Attn}(\mathbf{s}_q, \mathbf{s}_c) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \quad (6)$$

The resulting attention output is added to the query features via a residual connection, reshaped back to the spatial image format, and concatenated with the unchanged condition features.

4.2 Camera pose encoding

When a human observer struggles to separate reflections from transmissions in a complex scene, they naturally shift their viewpoint. By comparing their own movement against the apparent motion of objects in the scene, they can effortlessly distinguish the real background (transmission) from virtual images (reflection) through depth-induced parallax.

Inspired by this, we introduce relative camera pose encoding into our dual-view framework. This encoding serves as a global geometric prior, informing the network of the exact camera motion and guiding spatial attention to features that adhere to the transmission parallax, while suppressing reflection features whose apparent motion violates this transformation.

Given poses $\mathbf{p}_1 = [\mathbf{t}_1, \mathbf{q}_1]^\top$ and $\mathbf{p}_2 = [\mathbf{t}_2, \mathbf{q}_2]^\top$, where $\mathbf{t}_i \in \mathbb{R}^3$ represents translation and $\mathbf{q}_i \in \mathbb{R}^4$ is the unit quaternion for rotation, we compute the relative pose:

$$\mathbf{p}_{\text{rel}} = \left[\frac{\mathbf{t}_2 - \mathbf{t}_1}{\|\mathbf{t}_2 - \mathbf{t}_1\|}, \mathbf{q}_2 \otimes \mathbf{q}_1^{-1} \right]. \quad (7)$$

We normalize the translation vector to a unit direction because IMU-derived translations suffer from magnitude drift, whereas rotations and translation directions remain highly reliable.

To integrate this 7-dimensional relative pose, we process it through a lightweight 2-layer MLP (hidden dimension 256) to produce the pose embedding $\mathbf{e}_p = \text{MLP}(\mathbf{p}_{\text{rel}})$. This is then fused with the time embedding $\mathbf{e}_t$ to form the combined conditioning signal $\mathbf{e}_c = \mathbf{e}_t + \gamma\mathbf{e}_p$. We empirically set the scaling factor $\gamma = 0.15$ to prevent the pose information from overwhelming the temporal diffusion signal while maintaining sufficient geometric guidance.

Our approach fundamentally differs from prior burst-based methods [9] that rely on explicit 3D projections during test-time optimization. Because they enforce strict geometric constraints, these methods are computationally expensive, highly sensitive to noise, and require high-precision IMUs. In contrast, we inject camera pose as a soft conditional embedding within an end-to-end diffusion pipeline (Figure 4). This formulation provides robust geometric guidance that effectively tolerates the inherent drift and noise of common smartphone sensors.

4.3 Training strategy

We discover that training exclusively on simple reflection scenarios produces models unable to handle strong reflections, such as overexposed or structurally intense reflections. Conversely, directly training on challenging cases prevents the model from learning to leverage parallax information for understanding the physical principles underlying reflection-transmission decomposition. We therefore adopt two-stage curriculum learning that progressively increases difficulty.

Stage 1: Foundation training. We begin with semi-synthetic data at 768×768 resolution with moderate reflection intensity $\alpha_r \sim \mathcal{U}(0.3, 0.7)$. Crucially, we apply stochastic augmentations to enhance model robustness: with 15% probability, we replace view 2 with view 1’s content, and with another 15% probability, we replace view 1 with view 2’s content. In both cases, the relative pose is set to the identity (zero) since the views become identical. Additionally, when both views are distinct (in 70% of cases), we apply pose dropout with a 25% conditional probability, removing pose information to encourage the model to rely on visual features as a fallback. These augmentations enable the model to handle both single-view and dual-view scenarios gracefully while establishing core capabilities for reflection-transmission decomposition.

Stage 2: Robustness enhancement. We increase reflection intensity $\alpha_r \sim \mathcal{N}(0.8, 0.2)$ clipped to $[0.4, 1.2]$ and introduce diverse augmentations including motion blur, chromatic aberration, and ghosting effects. This stage fine-tunes the model for challenging scenarios where reflections may dominate transmissions. The training objective is the standard DDPM loss [20] with conditional dropout:

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{z}_0, \epsilon, m_v, m_p} [\|\epsilon - \epsilon_\theta(\mathbf{z}_T^t, t, m_v \odot \mathbf{c}_M, m_p \cdot \mathbf{e}_{\text{pose}})\|_2^2], \quad (8)$$

where $m_v \in \{0, 1\}^2$ and $m_p \in \{0, 1\}$ are dropout masks for views and poses.

5 Experiments

5.1 Experimental setup

Datasets. We evaluate our method quantitatively on three datasets, namely (1) our DUALREF-V: 195 real-world sequences featuring challenging scenarios, including museums, storefronts, and vehicles; (2) DUALREF-R test split: 2819 semi-synthetic pairs (2014 video-based synthesis, 805 physics-based rendering) with precise ground truth; (3) Niklaus dataset [46] test split: 20 real-captured dual-view pairs. Our model is trained on the DUALREF-R training split comprising 13,181 pairs. During evaluation, we sample 500 dual-view pairs from 195 captured sequences to approximately form a Gaussian distribution with a mean view shift of 50 pixels and a variance of 40, in accordance with the view distance of common dual-camera systems. Furthermore, we evaluate our method on images from various sources (the Chugunov dataset [9], in-the-wild Internet data, and real-captured photos) with diverse viewpoint distances.

Baselines. We compare against state-of-the-art single-image reflection removal methods including ZF24 [92], DSIT [30], RDNet [90], DAI [27], WSeal [87], and

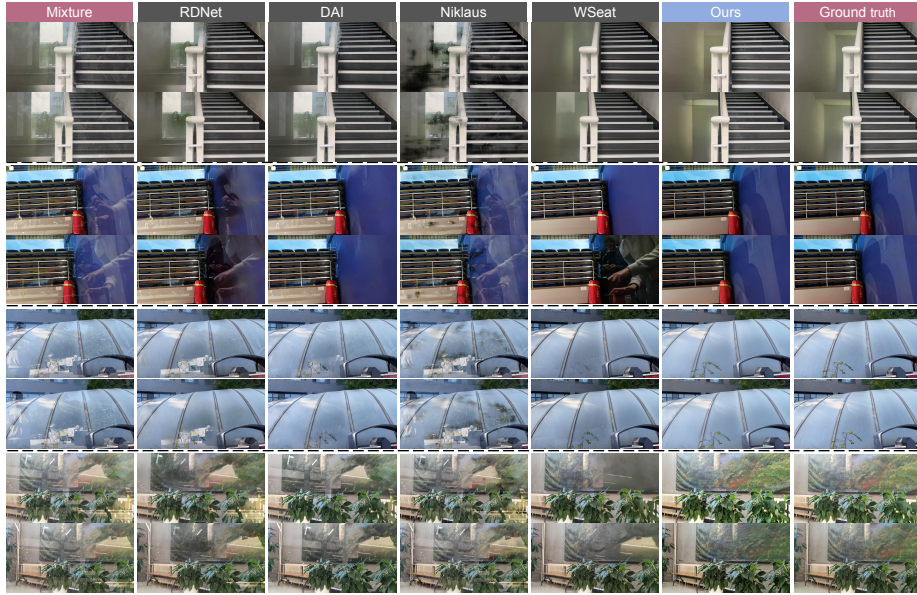


Fig. 5: We compare with RDNet [90], DAI [27], Niklaus [46] and Weat [87] on challenging real-world scenarios in our DUALREF-V dataset. For each scene divided by dashed lines, the upper row shows view 1, and the bottom row shows view 2. Readers are encouraged to zoom in for best viewing.

multi-view methods TF15 [81], Niklaus [46]. For a fair comparison, we also evaluate fine-tuned versions (marked with *) of methods on our training dataset (both views are separately used during finetuning of SIRR methods). L-Differ [26], Kee *et al.* [33], and Prasad *et al.* [50] are not open-source, yet we compare with Adobe Photoshop 2026, which is linked to Kee *et al.* [33], along with other general-purpose image-editing models [15, 47] in our supplementary materials.

Metrics. We employ five metrics: SSIM and PSNR for structural similarity and signal quality; LPIPS [88] for perceptual distance; CLIPQA [67] and MUSIQ [32] for no-reference image quality assessment. These metrics collectively evaluate pixel accuracy and perceptual quality.

5.2 Comparison to state-of-the-arts

As shown in Table 1, our method achieves superior performance across most evaluation metrics, especially on real-world data (DUALREF-V and the Niklaus [46] dataset), demonstrating the effectiveness of leveraging dual-view information through our diffusion framework.

As shown in Figure 5 and Figure 6, single-image methods [27, 30, 87, 90] often fail when confronted with overexposed or structurally complex reflections, sometimes even misidentifying reflections as transmission and enhancing them rather than removing them. Diffusion-based SIRR approaches like DAI [27] and



Fig. 6: Qualitative comparison on challenging real-world scenarios collected from diversified sources, including the Chugunov dataset [9], the Niklaus dataset [46], and videos downloaded from the Internet. For each scene separated by dashed lines, the upper row (or left column) and the bottom row (or right column) correspond to views 1 and 2, respectively. Our dual-view diffusion framework effectively removes complex and spatially varying reflections while faithfully preserving fine transmission details and scene structures. The images are best viewed zoomed in for a clearer comparison of local details. Please refer to our supplementary material for more results.

Table 1: Quantitative comparison on DUALREF and the Niklaus dataset [46]. Our method demonstrates superior performance across both real-world and semi-synthetic scenarios. * indicates methods fine-tuned on our training data. **Bold** and underlined values denote best and second-best results.

Dataset	Metric	TF15 [81]	ZF24 [92]	ZF24* DSIT [30]	DSIT* RDNet [90]	DAI [27]	WSeat [87]	Niklaus* [46]	Ours
DUALREF-V	SSIM↑	0.6854	0.7300	0.7317	0.7285	<u>0.7481</u>	0.7342	0.7354	0.7478
	PSNR↑	18.65	19.80	20.17	19.54	20.55	20.24	<u>21.57</u>	21.44
	LPIPS↓	0.2243	0.1923	0.2005	0.2072	0.1975	0.2014	0.1762	<u>0.1692</u>
	CLIPQA↑	0.3762	0.4420	0.4538	0.4550	0.4527	0.4569	0.4431	<u>0.4691</u>
	MUSIQ↑	61.77	63.53	64.01	63.56	63.52	<u>64.22</u>	62.89	64.00
DUALREF-R	SSIM↑	0.7868	0.8273	0.8359	0.8337	0.8389	0.8362	0.8274	0.8369
	PSNR↑	23.75	25.64	26.61	26.43	26.71	26.60	26.66	<u>27.11</u>
	LPIPS↓	0.1523	0.1133	0.1045	0.1003	<u>0.0990</u>	0.1017	0.1102	0.1448
	CLIPQA↑	0.3425	0.3778	0.3690	0.3575	0.3705	0.3826	<u>0.4024</u>	0.3981
	MUSIQ↑	44.32	46.26	45.94	47.07	46.85	<u>48.25</u>	47.11	47.90
Niklaus [46]	SSIM↑	0.7575	0.7726	0.7743	0.7999	0.7971	0.7969	0.7842	0.7893
	PSNR↑	21.76	22.61	22.88	22.71	23.40	23.27	23.33	<u>23.72</u>
	LPIPS↓	0.1432	0.1209	0.1185	0.1168	0.1154	<u>0.0934</u>	0.1813	0.1669
	CLIPQA↑	0.4882	0.4952	0.4976	0.4676	0.4712	0.5989	0.3461	0.5909
	MUSIQ↑	51.94	52.37	52.66	53.29	53.57	52.71	50.81	56.31

Table 2: Ablation study on core components. The first and second set of experiments ablate the pose encoding and SD2.1’s generative prior, respectively. The third set of experiment compares different attention mechanisms.

Configuration	DUALREF-V					DUALREF-R				
	PSNR↑	SSIM↑	LPIPS↓	CLIPQA↑	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	CLIPQA↑	MUSIQ↑
Single-view	21.59	0.7236	0.1720	0.4571	62.13	26.07	0.8235	0.1280	0.4096	45.27
Dual separate	21.44	0.7245	0.1703	0.4597	63.77	26.12	0.8219	0.1248	0.4231	45.12
Dual w/o pose	23.53	0.7468	0.1694	0.4632	63.89	27.47	0.8310	0.1176	0.4270	47.26
SD2.1+RAFT	21.60	0.7369	0.1744	0.4588	63.21	26.33	0.8226	0.1265	0.4178	45.19
Cat3D-Style Attn	22.09	0.7267	0.1717	0.4588	62.40	26.52	0.8230	0.1244	0.4133	45.68
SIR-Style Attn	20.87	0.7036	0.2069	0.3678	55.32	25.99	0.7963	0.1452	0.3685	43.20
Width-concat	22.03	0.7354	0.1838	0.4234	58.79	25.83	0.8220	0.1392	0.4002	42.99
Ours	24.17	0.7591	0.1638	0.4705	64.81	28.19	0.8378	0.0983	0.4384	48.76

WindowSeat [87] tend to alter content structure. In contrast, our method robustly handles these challenging scenarios without such artifacts while maintaining user-friendliness by using dual-camera systems that require only a single shutter press. As shown in Figure 5, optical flow-based dual-view methods [46,81] struggle with strong reflections, which hinder accurate flow estimation, leading to error propagation and unsatisfactory results. In contrast, our approach circumvents these issues. Furthermore, our method generalizes well to challenging in-the-wild scenes with diverse cross-view distances, which robustly adapts to various dual-camera devices.

5.3 Ablation study

As shown in Table 2, Figure 7 and Figure 8, we conduct comprehensive ablation studies using models trained on the DUALREF-R train split and evaluated on both DUALREF-R test split and DUALREF-V. Our analysis is twofold: verifying the core components and validating our specific architectural design.

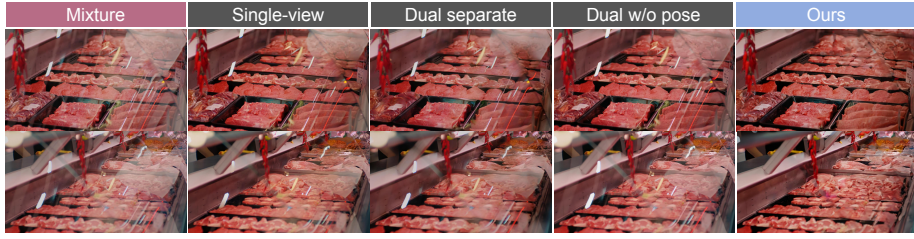


Fig. 7: Visual results for our ablation study. The upper row and the bottom row display views 1 and 2, respectively. Single-view denotes directly fine-tuning InstructPix2Pix [3] on DUALREF-R dataset (considering dual-view pairs as two separate single-views). Dual separate denotes the input of two identical images. Dual w/o pose denotes removing the pose encoding from our method.



Fig. 8: Analysis of explicit flow on the Burst dataset [9]. We compare against the SD2.1+RAFT and TF15 [81] baseline with different number of input frames. Explicit flow becomes unreliable in reflection-corrupted regions and needs more frames to reduce ambiguity, whereas our method produces robust results from only two views.

Analysis of pose encoding. The Single-view baseline (fine-tuned InstructPix2Pix without cross-view attention) and Dual separate (identical inputs with zero relative pose) both achieve strong baseline performance, proving that our base diffusion pipeline is inherently well-suited for reflection removal. Upgrading to Dual w/o pose brings clear quantitative gains, confirming that the network effectively leverages dual-view parallax to suppress reflections. Finally, our full model (**Dual w/ pose**) yields the highest performance across all metrics. As shown in Figure 7, Dual w/o pose fails to suppress intense reflections when they also demonstrate distinct parallax, as they become easily confused with the physical transmission. However, by informing the model of the actual viewpoint motion via our pose encoding integration scheme, the reflections are robustly separated through the correlation between camera trajectory and scene shift. This clearly validates our design choice. Readers are referred to the supplementary material for evaluations of other pose integration methods.

Analysis of explicit flow. To isolate the impact of SD2.1’s pre-trained generative prior, we built an SD2.1+RAFT baseline using identical training settings, using RAFT [61] to warp view 2 to view 1 before feeding the channel-wise concatenated latents into the U-Net, which consistently underperforms ours, as shown in Table 2. While explicit flow assumes a unique motion per pixel, reflection-degraded regions exhibit dual motions (transmission and reflection),

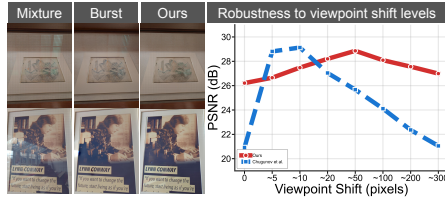


Fig. 9: Left: Comparison with state-of-the-art burst method [9] on their dataset. Our method achieves comparable or better quality while running $60\times$ faster. Right: Performance across different viewpoint shift levels on DUALREF-R test set, demonstrating our method’s robustness compared to burst approaches.

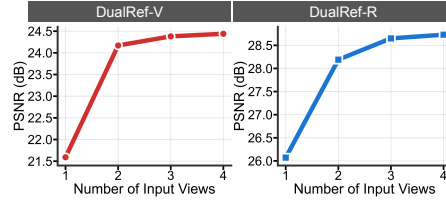


Fig. 10: Performance analysis with varying numbers of input views. We observe substantial improvements from single-view to dual-view, whereas additional views offer diminishing returns with minimal marginal gains. Our dual-view setup achieves an optimal balance between quality gains and capture complexity.

creating fundamental correspondence ambiguity. Our frame-aware attention bypasses this by implicitly aggregating evidence from multiple plausible matches without explicit commitment. Furthermore, flow-based methods like TF15 [81] require multiple frames to resolve such ambiguity and are particularly error-prone in the dual-view regime (Figure 8), where ours remains robust.

Analysis of cross-view attention mechanisms. We compare our frame-aware spatial attention against recent multi-view diffusion paradigms. The Cat3D-style baseline [14, 56] (naive channel-wise concatenation) suffers from cross-view feature ambiguity and noise leakage, visually manifesting as incomplete reflection removal and suboptimal quantitative metrics. Even worse, the SIR-style approach [45] (incorporating 3D ResNet blocks) aggressively alters the network topology, thereby destroying the delicate pre-trained 2D generative priors. Width-concat [78] converges slower and yields inferior results (Table 2). Unlike [78], inputs are inherently pixel-aligned in our task, where channel-wise fusion better exploits this direct correspondence. In contrast, our zero-overhead, directed attention perfectly preserves image fidelity while maximizing cross-view disambiguation. We also conduct ablation studies on the Niklaus dataset [46] with Dual w/o pose mode since pose data is not provided in this dataset. Please see our supplementary materials.

5.4 Discussions

Robustness to different levels of viewpoint shift. Figure 9 compares our method with the state-of-the-art burst approach [9]. The visual comparison (left) shows that our method achieves comparable or superior quality on their dataset while being significantly more efficient—requiring only 3 seconds per scene on an RTX 6000, compared to 3 minutes for burst methods, which require scene-specific optimization. The quantitative analysis (right) evaluates performance across varying viewpoint shifts on DUALREF-R test set. While burst methods

perform well at minimal displacements but degrade rapidly with larger shifts or single-view inputs, our approach maintains robust performance across all displacement levels, validating its robustness to different dual cameras with varying view distances. This robustness stems from our diverse training data and the synergy between frame-aware spatial attention and pose encoding.

Effectiveness of dual-view setup. We investigate the optimal number of views for reflection removal by training models with different numbers of input views on our dataset. Figure 10 reveals an interesting pattern: performance improves substantially from single-view to dual-view, but additional views beyond two provide diminishing returns. The marginal improvement from 2 to 4 views is minimal, while the capture complexity and inference time increase linearly. This validates our dual-view design as a reasonable balance—it achieves most of the quality gains possible from multi-view information while maintaining practical usability. Users need only a single shutter press with dual-camera systems, which is common in modern mobile devices, avoiding the operational burden of multi-frame sequences while still achieving significant improvements in reflection removal over single-image methods.

6 Conclusion

We present a dual-view diffusion framework for reflection removal that addresses the fundamental ill-posedness of single-image decomposition through frame-aware spatial attention and camera pose encoding. With dual-camera systems common in current devices, our approach provides a reasonable balance between capture simplicity and removal quality—requiring only a single shutter press with ubiquitous dual-camera systems on mobile device while achieving substantial improvements over single-image methods. We introduce DUALREF, the first comprehensive dual-view dataset with trajectory-aligned dual-view real-world sequences. Extensive experiments demonstrate superior performance across diverse scenarios, validating dual-view reflection removal as a practical paradigm for mobile photography.

Limitations. Our approach relies on effective parallax. In scenarios where view-point shift yields no or negligible parallax and insufficient geometric cues to disambiguate layer depths, the task reverts to an ill-posed single-image problem, and our performance naturally degrades to match single-image diffusion baselines. As a diffusion-based method, our model may also exhibit color drift in some challenging cases, although the dual-view cue can partially mitigate this issue by providing additional cross-view evidence. We further discuss our failure modes in our supplementary materials.

Acknowledgement. This work is supported by National Natural Science Foundation of China (Grant No. 62136001), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012).

References

1. Alayrac, J.B., Carreira, J., Zisserman, A.: The visual centrifuge: Model-free layered video representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2457–2466 (2019)
2. Apple Inc.: Live photos. <https://developer.apple.com/design/human-interface-guidelines/live-photos> (2024), accessed: 2025-11-11
3. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
4. Cai, J., Yang, K., Ouyang, L., Fu, L., Ding, J., Shen, J., Meng, Z.: OpenRR-5K: A large-scale benchmark for reflection removal in the wild. arXiv preprint arXiv:2506.05482 (2025)
5. Chang, Y., Jung, C., Sun, J.: Joint reflection removal and depth estimation from a single image. *IEEE Transactions on Cybernetics* **51**(12), 5836–5849 (2020)
6. Chang, Z., Weng, S., Li, Y., Li, S., Shi, B.: L-CoDer: Language-based colorization with color-object decoupling transformer. In: European Conference on Computer Vision. pp. 360–375. Springer (2022)
7. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-CoIns: Language-based colorization with instance awareness. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19221–19230 (2023)
8. Chaturvedi, S., Ren, M., Hold-Geoffroy, Y., Liu, J., Dorsey, J., Shu, Z.: Synthlight: Portrait relighting with diffusion model by learning to re-render synthetic faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 369–379 (2025)
9. Chugunov, I., Shustin, D., Yan, R., Lei, C., Heide, F.: Neural spline fields for burst image fusion and layer separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25763–25773 (2024)
10. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems* **34**, 8780–8794 (2021)
11. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5017–5026 (2021)
12. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3238–3247 (2017)
13. Fang, S., Wang, Y., Zhang, J., Li, Z., Wang, Y.: Adaptive language-aware image reflection removal network. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 973–981 (2025). <https://doi.org/10.24963/ijcai.2025/109>
14. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3D: Create anything in 3D with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
15. Google DeepMind: Gemini 3 pro image. <https://deepmind.google/models/gemini-image/pro/> (2026), accessed: 2026-02-27
16. Guo, X., Cao, X., Ma, Y.: Robust separation of reflection from multiple images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2187–2194 (2014)
17. Guo, Y., Lao, Z., Song, X., Zhou, Y., Yu, H.: Sirr-lmm: Single-image reflection removal via large multimodal model (2026), <https://arxiv.org/abs/2601.07209>

18. Han, B.J., Sim, J.Y.: Reflection removal using low-rank matrix completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5438–5446 (2017)
19. Han, B.J., Sim, J.Y.: Zero-shot learning for reflection removal of single 360-degree image. In: Proceedings of the European Conference on Computer Vision. pp. 533–548 (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
21. Hong, Y., Chang, Y., Liang, J., Ma, L., Huang, T., Shi, B.: Light flickering guided reflection removal. *International Journal of Computer Vision* **132**(9), 3933–3953 (2024)
22. Hong, Y., Lyu, Y., Li, S., Cao, G., Shi, B.: Reflection removal with NIR and RGB image feature fusion. *IEEE Transactions on Multimedia* **25**, 7101–7112 (2022)
23. Hong, Y., Lyu, Y., Li, S., Shi, B.: Near-infrared image guided reflection removal. In: IEEE International Conference on Multimedia and Expo. pp. 1–6. IEEE (2020)
24. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: Panoramic image reflection removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7762–7771 (2021)
25. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: Par2Net: End-to-end panoramic image reflection removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(10), 12192–12205 (2023)
26. Hong, Y., Zhong, H., Weng, S., Liang, J., Shi, B.: L-Differ: Single image reflection removal with language-based diffusion model. In: Proceedings of the European Conference on Computer Vision. pp. 58–76 (2024)
27. Hu, J., Yang, C., Zhou, Z., Fang, J., Yang, X., Tian, Q., Shen, W.: Dereflexion any image with diffusion priors and diversified data. *arXiv preprint arXiv:2503.17347* (2025)
28. Hu, Q., Guo, X.: Trash or Treasure? An interactive dual-stream strategy for single image reflection separation. *Advances in Neural Information Processing Systems* **34**, 24683–24694 (2021)
29. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13138–13147 (2023)
30. Hu, Q., Wang, H., Guo, X.: Single image reflection separation via dual-stream interactive transformers. *Advances in Neural Information Processing Systems* **37**, 55228–55248 (2024)
31. Kavar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. *Advances in Neural Information Processing Systems* **35**, 23593–23606 (2022)
32. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
33. Kee, E., Pikielny, A., Blackburn-Matzen, K., Levoy, M.: Removing reflections from RAW photos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 161–171. Computer Vision Foundation / IEEE (2025)
34. Kong, N., Tai, Y.W., Shin, S.Y.: A physically-based approach to reflection separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9–16. IEEE (2012)
35. Lei, C., Chen, Q.: Robust reflection removal with reflection-free flash-only cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14811–14820 (2021)

36. Lei, C., Huang, X., Qi, C., Zhao, Y., Sun, W., Yan, Q., Chen, Q.: A categorized reflection removal dataset with diverse real-world scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3040–3048 (2022)
37. Lei, C., Huang, X., Zhang, M., Yan, Q., Sun, W., Chen, Q.: Polarized reflection removal with perfect alignment in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1750–1758 (2020)
38. Lei, C., Jiang, X., Chen, Q.: Robust reflection removal with flash-only cues in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(12), 15530–15545 (2023)
39. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3565–3574 (2020)
40. Li, X., Wang, Z., Huang, Z., Rehg, J.: Cue3D: Quantifying the role of image cues in single-image 3D generation. *Advances in Neural Information Processing Systems* **38**, 105199–105232 (2026)
41. Liang, Z., Yu, H., Li, B., Zhang, D., Cao, Z., Gong, T., Liu, J., Cui, S., Wang, F.: 3DReflecNet: A large-scale dataset for 3D reconstruction of reflective, transparent, and low-texture objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7244–7255 (2026)
42. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Re-Paint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
43. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Reflection separation using a pair of unpolarized and polarized images. *Advances in Neural Information Processing Systems* **32** (2019)
44. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Physics-guided reflection separation from a pair of unpolarized and polarized images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(2), 2151–2165 (2022)
45. Mao, Y., Wang, B., Kulkarni, N., Park, J.J.: SIR-DIFF: Sparse image sets restoration with multi-view diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21620–21630 (2025)
46. Niklaus, S., Zhang, X.C., Barron, J.T., Wadhwa, N., Garg, R., Liu, F., Xue, T.: Learned dual-view reflection removal. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3713–3722 (2021)
47. OpenAI: Openai chatgpt 5.2. <https://chatgpt.com/> (2025), accessed: 2026-02-27
48. Park, J., Kim, H., Park, E., Sim, J.Y.: Fully-automatic reflection removal for 360-degree images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1609–1617 (2024)
49. Poly Haven: Poly haven. <https://polyhaven.com/> (2026), accessed: 2026-06-22
50. Prasad, B., Mitra, K., et al.: Burst reflection removal using reflection motion aggregation cues. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 239–248 (2023)
51. Punnapurath, A., Brown, M.S.: Reflection removal using a dual-pixel sensor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1556–1565 (2019)
52. Qiu, J., Jiang, P.T., Zhu, Y., Yin, Z.X., Cheng, M.M., Ren, B.: Looking through the glass: Neural surface reconstruction against high specular reflections. In: Proceed-

- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20823–20833 (2023)
53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
 54. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
 55. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(4), 4713–4726 (2022)
 56. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: MVDream: Multi-view diffusion for 3D generation. *arXiv preprint arXiv:2308.16512* (2023)
 57. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3193–3201 (2015)
 58. Simon, C., Kyu Park, I.: Reflection removal for in-vehicle black box videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4231–4239 (2015)
 59. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020)
 60. Song, Z., Zhang, Z., Zhang, K., Luo, W., Fan, Z., Ren, W., Lu, J.: Robust single image reflection removal against adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24688–24698 (2023)
 61. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: Proceedings of the European Conference on Computer Vision. pp. 402–419 (2020)
 62. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3922–3930 (2017)
 63. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Face image reflection removal. *International Journal of Computer Vision* **129**(2), 385–399 (2021)
 64. Wan, R., Shi, B., Li, H., Duan, L.Y., Tan, A.H., Kot, A.C.: CoRRN: Cooperative reflection removal network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(12), 2969–2982 (2019)
 65. Wan, R., Shi, B., Li, H., Hong, Y., Duan, L.Y., Kot, A.C.: Benchmarking single-image reflection removal algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(2), 1424–1441 (2022)
 66. Wang, C., Xu, D., Wan, R., He, B., Shi, B., Duan, L.Y.: Background scene recovery from an image looking through colored glass. *IEEE Transactions on Multimedia* **25**, 2876–2887 (2022)
 67. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
 68. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision* **132**(12), 5929–5949 (2024)
 69. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision* **132**(12), 5929–5949 (2024)

70. Wang, T., Lu, W., Zhang, K., Luo, W., Kim, T.K., Lu, T., Li, H., Yang, M.H.: PromptRR: Diffusion models as prompt generators for single image reflection removal. arXiv preprint arXiv:2402.02374 (2024)
71. Wang, T., Xie, M., Cai, H., Shah, S., Metzler, C.A.: Flash-Split: 2D reflection removal with flash cues and latent diffusion separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5688–5698 (2025)
72. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: SinSR: Diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25796–25805 (2024)
73. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8178–8187 (2019)
74. Wen, Q., Tan, Y., Qin, J., Liu, W., Han, G., He, S.: Single image reflection removal beyond linearity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3771–3779 (2019)
75. Weng, S., Wu, H., Chang, Z., Tang, J., Li, S., Shi, B.: L-CoDe: Language-based colorization using color-object decoupled conditions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2677–2684 (2022)
76. Weng, S., Zhang, P., Li, Y., Li, S., Shi, B., et al.: L-CAD: Language-based colorization with any-level descriptions using diffusion priors. *Advances in Neural Information Processing Systems* **36**, 77174–77186 (2023)
77. Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A.G., Milanfar, P.: Deblurring via stochastic refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16293–16303 (2022)
78. Winter, D., Shul, A., Cohen, M., Berman, D., Pritch, Y., Rav-Acha, A., Hoshen, Y.: ObjectMate: A recurrence prior for object insertion and subject-driven generation. arXiv preprint arXiv:2412.08645 (2024)
79. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: DiffIR: Efficient diffusion model for image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13095–13105 (2023)
80. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: SmartBrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023)
81. Xue, T., Rubinstein, M., Liu, C., Freeman, W.T.: A computational approach for obstruction-free photography. *ACM Transactions on Graphics* **34**(4), 1–11 (2015)
82. Yang, J., Gong, D., Liu, L., Shi, Q.: Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In: Proceedings of the European Conference on Computer Vision. pp. 654–669 (2018)
83. Yang, K., Ouyang, L., Sun, H., Cai, J., Fu, L., Ding, J., Ho, C.M., Meng, Z.: OpenRR-1K: A scalable dataset for real-world reflection removal. arXiv preprint arXiv:2506.08299 (2025)
84. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
85. Yu, K., Han, M., Pan, L., Liu, L., Liu, M., Liang, W.: DCRR++: Unsupervised reflection removal and novel view synthesis via dual-pixel guided 3d gaussian splatting. *International Journal of Computer Vision* **134**(6), 284 (2026)

86. Yu, K., Pan, L., Liu, L., Liang, W.: Enhanced dual-pixel image reflection removal via gaussian splatting. In: Proceedings of the ACM International Conference on Multimedia. pp. 7766–7775 (2025)
87. Zakarin, D., Wandel, T., Obukhov, A., Dai, D.: Reflection removal through efficient adaptation of diffusion transformers. arXiv preprint arXiv:2512.05000 (2025)
88. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
89. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4786–4794 (2018)
90. Zhao, H., Li, M., Hu, Q., Guo, X.: Reversible decoupling network for single image reflection removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26430–26439 (2025)
91. Zhong, H., Hong, Y., Weng, S., Liang, J., Shi, B.: Language-guided image reflection separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24913–24922 (2024)
92. Zhu, Y., Fu, X., Jiang, P.T., Zhang, H., Sun, Q., Chen, J., Zha, Z.J., Li, B.: Revisiting single image reflection removal in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25468–25478 (2024)