PRISM: Latent Composition Consistency for Single-Image Reflection Removal

Junseong Shin[1] and Tae Hyun Kim[1,2⋆]

[1] Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
[2] Department of Computer Science, Hanyang University, Seoul, South Korea
{junsung6140, taehyunkim}@hanyang.ac.kr

Abstract. Single-image reflection removal (SIRR) seeks to recover the transmission layer from a mixture corrupted by reflections—a severely ill-posed problem. Existing methods operate in pixel space, where the nonlinear sRGB formation model entangles the two layers and limits generalization. We observe that pretrained VAE latent spaces exhibit substantially lower coherence between image layers compared to pixel space, providing a more favorable working space for decomposition. Building on this finding, we propose **PRISM** (Pretrained-latent Reflection Image Separation Model), which reinterprets SIRR as a latent linear separation problem. Under an approximate additive formulation in latent space, PRISM learns a flow matching velocity field on a pretrained FLUX backbone that recovers both transmission and reflection in a single forward pass. To enforce robust disentanglement, we introduce a Latent Composition Consistency (LCC) strategy that constructs synthetic mixtures by swapping reflection latents across samples and enforces consistent decomposition via a cycle loss. We further propose a Layer Contrastive Separation (LCS) loss that promotes semantic separation between layers through patch-level contrastive learning, without requiring explicit reflection targets. Experiments on six benchmarks demonstrate that PRISM consistently outperforms state-of-the-art methods by significant margins, with strong generalization to in-the-wild images.

Keywords: Reflection Removal · Flow Matching · Contrastive Learning

1 Introduction

Photographs taken through glass windows or other transparent surfaces routinely capture unwanted reflections superimposed on the scene of interest. Beyond degrading visual quality, such reflections harm downstream tasks including 3D Gaussian Splatting [37], face recognition [36], and depth estimation [2]. Single-image reflection removal (SIRR) seeks to recover the transmission layer T from an observed mixture $I = T + R$, where R is the reflection component [29]. This problem is severely ill-posed, as infinitely many (T, R) pairs satisfy the equation for a given I.

⋆ Corresponding author

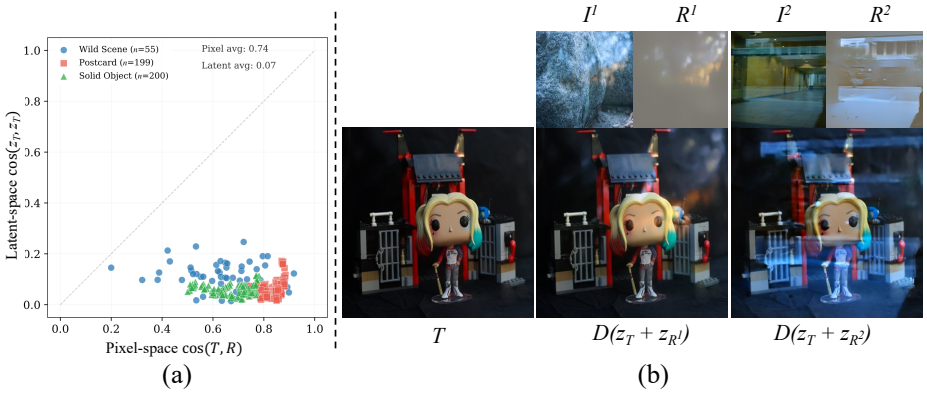

Fig. 1: (a) Cosine similarity between transmission and reflection in pixel space vs. FLUX VAE [18] latent space on 454 pairs from SIR^2 [35]. All points lie below the diagonal, confirming consistently lower coherence in latent space. **(b)** Latent swap results: composing $\hat{z}_T$ with reflections $\hat{z}_R^1$, $\hat{z}_R^2$ separated from different images and decoding via the VAE decoder $\mathcal{D}$ yields realistic mixtures that preserve transmission content while swapping only the reflection, demonstrating well-disentangled decomposition.

Prior works address this ambiguity through hand-crafted priors [16, 20, 33], learnable residue terms [13,48], or data-driven architectures [11,14,47]. However, operating in pixel space imposes two fundamental limitations: the additive model is only a coarse approximation under nonlinear camera responses, and architectures must simultaneously capture low-level texture and high-level semantics, often resulting in entangled features. While the RAW sensor domain offers a more faithful additive model [15], it requires specialized hardware inapplicable to existing sRGB imagery.

Why latent space? We pursue an orthogonal direction: lifting the decomposition into the latent space of a pretrained generative model [17, 18, 32]. As shown in Fig. 1(a), we measure the cosine similarity between transmission and reflection across 454 real image pairs from SIR^2 [35]. In pixel space, T and R are highly entangled (avg. cos = 0.74), whereas in the FLUX VAE latent space the cosine similarity drops to 0.07 — a reduction shared by any pair of encoded images, indicating broad decorrelation rather than a layer-specific property. This substantially lower coherence suggests that the latent space serves as a more favorable working space for decomposition. Leveraging this property, we adopt an approximate additive formulation $z_I \approx z_T + z_R$, where $z = \mathcal{E}(\cdot)$ denotes the output of a pretrained VAE encoder $\mathcal{E}$ and $\mathcal{D}(\cdot)$ its decoder. We model the transport $z_I \to z_T$ via a flow matching velocity field v_θ. The velocity target satisfies $v^* \approx -z_R$, enabling both components to be recovered in a single forward pass without dual-branch architectures.

We term our method **PRISM** (**P**retrained-latent **R**eflection **I**mage **S**eparation **M**odel). A key challenge is preventing degenerate solutions where the model entangles scene-specific information across components. The low mutual coherence

of latent representations offers a unique opportunity here: since the overlap between layers is substantially lower in latent space, composing the transmission of one image with the reflection of another yields a semantically valid mixture (Fig. 1(b)) — something far less effective in pixel space where the two layers overlap heavily. We exploit this property with a **Latent Composition Consistency (LCC)** strategy that swaps reflection components across images to construct synthetic latent compositions, enforcing that the model decomposes them consistently.

Since the additive formulation $z_I \approx z_T + z_R$ is only approximate, an explicit reflection regression target would require encoding a pixel-space residual ($\mathcal{E}(I-T)$), propagating nonlinearity errors into the training signal. We therefore introduce a **Layer Contrastive Separation (LCS)** loss that enforces semantic separation between transmission and reflection without requiring any explicit reflection target.

Our contributions are as follows:

- We provide empirical evidence that pretrained VAE latent spaces exhibit significantly lower coherence between image layers compared to pixel space, and show that this property enables effective latent-space decomposition.
- We propose **PRISM**, a latent flow matching framework that reinterprets SIRR as a latent linear separation problem, recovering both transmission and reflection in a single forward pass without dual-branch architectures.
- We introduce a **Latent Composition Consistency (LCC)** training strategy that leverages the low coherence of latent space to construct semantically valid cross-image compositions, enforcing decomposition robustness across arbitrary latent mixtures.
- We introduce a **Layer Contrastive Separation (LCS)** loss that enforces semantic separation between transmission and reflection without requiring exact additive reconstruction, directly addressing the non-exact nature of the latent additive assumption.

2 Related Works

2.1 Single-Image Reflection Removal

Early work models the observed image as $I = T + R$, where T and R denote the transmission and reflection layers [29]. Subsequent refinements incorporate ghosting cues [33], alpha-matting coefficients [20], channel-wise blending scalars [16], and a learnable residue term $\Phi(T, R)$ to capture nonlinear interactions [13]. Operating in the RAW sensor domain makes this additive model more physically faithful [15], but requires specialized hardware unavailable for standard sRGB imagery.

With paired datasets, supervised networks have become dominant. Single-stream methods such as CEILNet [8], Zhang *et al.* [46], and ERRNet [39] learn direct mappings to the transmission layer via task-specific architectures and

losses. Recognizing the complementary nature of the two layers, dual-stream architectures model both jointly: IBCLN [19] uses a recurrent dual-branch LSTM; YTMT [12] introduces a symmetric ReLU-based interaction strategy; LASIRR [5] adds explicit reflection localization with confidence maps; and [48], DSRNet [13], and DSIT [14] progressively advance feature interaction quality via gated modules, learnable residue terms, and dual-attention mechanisms, respectively. RDNet [47] further addresses information loss in dual-stream interactions through multi-level reversible encoders, while Zhu *et al.* [50] contribute a real-world dataset to mitigate data scarcity. More recently, diffusion-based priors [10] have been brought to bear on SIRR: DAI [11] demonstrates that constraining predictions to a natural-image manifold substantially improves real-world generalization, though it requires fine-tuning the Stable Diffusion [32] VAE to compensate for its limited reconstruction fidelity. *Despite this progress, all of the above methods operate in pixel space and are susceptible to the nonlinear sRGB formation model; none enforces a structured additive decomposition in the representation space of a generative model.*

2.2 Flow Matching and Latent Generative Models

VAEs [17] and latent diffusion models [32] establish that compressed latent spaces are smooth, semantically structured, and amenable to compositional arithmetic—properties exploited for flexible image editing [27] and generation. Flow matching [22] learns a velocity field transporting samples along straight trajectories; Rectified Flow [23] further straightens these paths, and consistency models [34] push this to single-step generation. FLUX [18] instantiates these principles at scale, providing a pretrained latent space well-suited for task-specific fine-tuning. Generative priors have also been applied to image restoration [10], with diffusion- and flow-based methods [21, 26, 43, 49] showing that constraining predictions to a natural-image manifold substantially improves real-world generalization. *Unlike these methods, which repurpose generative models as pixel-space priors or require iterative sampling, we directly exploit the additive structure of the latent space for single-pass reflection decomposition.*

2.3 Contrastive Learning for Image Restoration

Contrastive learning [3, 9, 30] trains representations by pulling positive pairs together while pushing negatives apart, and has been increasingly adopted in low-level vision tasks. AECR-Net [42] first introduced contrastive regularization for image dehazing by treating hazy images as negatives and their clean counterparts as positives, establishing the paradigm of using degradation-clean pairs as contrastive supervision. CUT [31] further demonstrated that *patch-wise* contrastive objectives can effectively align local structures in unpaired image-to-image translation, inspiring subsequent patch-level formulations. More recently, task-agnostic contrastive strategies [41] have been proposed to improve restoration quality across diverse degradation types. *Unlike these works, which apply contrastive objectives to improve single-target reconstruction quality, our Layer*

Contrastive Separation loss repurposes this framework for cross-layer disentanglement between transmission and reflection.

3 Method

3.1 Preliminary: Latent Flow Matching and FLUX

Flow Matching. Flow matching [22] learns a velocity field v_θ that transports samples from a source distribution p_0 to a target distribution p_1 along straight trajectories. Given a sample pair (x_0, x_1), the conditional flow is defined as the linear interpolation $x_t = (1-t)x_0 + tx_1$ for $t \in [0,1]$, with constant velocity target $v^* = x_1 - x_0$. At inference, the one-step estimate recovers the target as $x_1 = x_0 + v_\theta(x_0, 0)$.

Latent Flow Matching. Rather than operating in pixel space, latent flow matching [6, 40] performs this transport in the compressed representation of a pretrained VAE [17]. An encoder $\mathcal{E}$ maps an image I to a latent $z = \mathcal{E}(I) \in \mathbb{R}^{16\times H/8\times W/8}$ (16 channels, 8× spatial downsampling) and a decoder $\mathcal{D}$ inverts this mapping. The latent space is semantically structured and exhibits smooth compositional properties [32], making it well-suited for the additive decomposition we exploit in Sec. 3.2. Crucially, FLUX's VAE achieves high reconstruction fidelity without additional fine-tuning, unlike Stable Diffusion-based approaches [11, 32] that require VAE fine-tuning to compensate for reconstruction artifacts.

FLUX. FLUX [18] is a large-scale instantiation of latent flow matching built on a Multi-Modal Diffusion Transformer (MM-DiT) backbone [6]. We adopt FLUX as the backbone for our reflection removal framework, as detailed in Sec. 3.3. Conventional diffusion-based image-to-image methods begin from a noisy latent obtained by adding noise to the encoded input, and iteratively denoise toward the target. FLUX, as a flow matching model, supports starting directly from a clean input latent without any noise injection, since the straight-line trajectory allows recovery at any timestep. We exploit this property: PRISM takes the clean input latent $z_I = \mathcal{E}(I)$ as the starting point and recovers the transmission latent in a single forward pass.

3.2 Problem Formulation

We adopt the standard additive image formation model [29, 35, 46], where the observed image I is expressed as the superposition of a transmission layer T and a reflection layer R:

$$I = T + R. \tag{1}$$

This formulation serves only as an approximation under real imaging conditions due to nonlinear camera responses, blur, and attenuation effects. Instead of refining the physical model in pixel space, we perform decomposition in the VAE

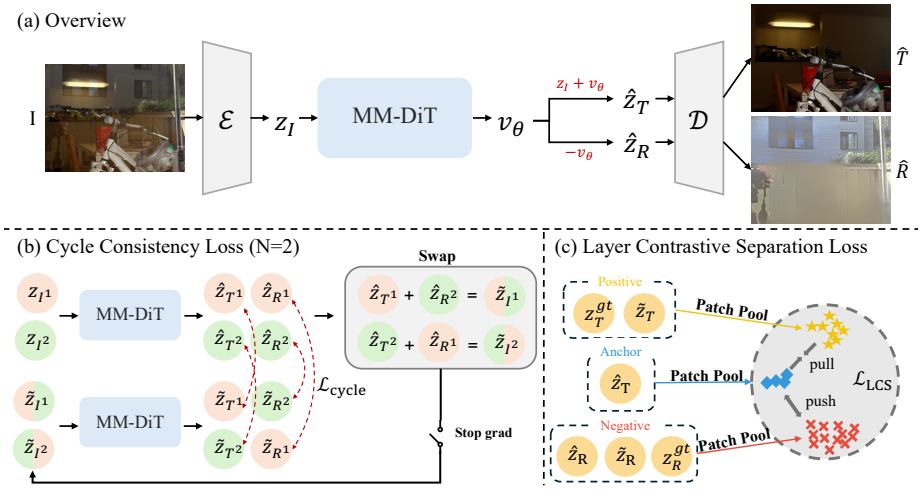

Fig. 2: Overview of PRISM. (a) MM-DiT predicts v_θ from $z_I = \mathcal{E}(I)$ (VAE encoder $\mathcal{E}$, decoder $\mathcal{D}$), yielding $\hat{z}_T = z_I + v_\theta$ and $\hat{z}_R = -v_\theta$ in a single pass; both are decoded by $\mathcal{D}$ to produce $\hat{T}$ and $\hat{R}$. **(b)** Reflection latents are cyclically swapped across batch samples to form synthetic mixtures $\tilde{z}_I$; a second forward pass enforces consistent decomposition ($\mathcal{L}_{\text{cycle}}$). **(c)** Patch-pooled features of the cycle-recovered transmission $\tilde{z}_T$ are pulled toward positive (transmission) and pushed from negative (reflection) samples via InfoNCE ($\mathcal{L}_{\text{LCS}}$).

latent space introduced above, where encoded representations exhibit substantially lower mutual coherence (Fig. 1(a)). We adopt an approximate additive formulation:

$$z_I \approx z_T + z_R, \tag{2}$$

where $z_I = \mathcal{E}(I)$, $z_T = \mathcal{E}(T)$, and $z_R = \mathcal{E}(R)$. Although this constraint does not hold exactly due to encoder nonlinearity, the low coherence in latent space eases the regression and makes it a practically effective working assumption, casting SIRR as a *latent linear separation* problem.

3.3 Latent Flow Matching for Reflection Removal

Given the latent additive formulation $z_I \approx z_T + z_R$, our goal is to estimate z_T from z_I alone. We model this transformation using a flow matching framework, where a velocity field v_θ is learned to transport the input latent z_I toward the transmission latent z_T. Specifically, we define the velocity target as:

$$v^* = z_T - z_I \approx -z_R, \tag{3}$$

which reveals a key property: under the additive latent assumption, the velocity field approximates the negative of the reflection latent, allowing both components

to be recovered in a single forward pass without any dual-branch architecture or dedicated reflection loss (Fig. 2(a)):

$$\hat{z}_T = z_I + v_\theta(z_I, 0), \qquad \hat{z}_R = -v_\theta(z_I, 0). \tag{4}$$

We implement v_θ by fine-tuning the full parameters of a pretrained FLUX model [18], with z_I serving directly as the starting point of the flow. Rather than sampling $t \sim \mathcal{U}[0, 1]$ during training, we fix $t = 0$ and train with a single-step objective directly on the input latent z_I, which reduces training cost and is consistent with one-step inference:

$$\mathcal{L}_{\text{latent}} = \|v_\theta(z_I, 0) - v^*\|_2^2. \tag{5}$$

To further improve perceptual quality, we additionally supervise the decoded transmission prediction in pixel space using a combination of ℓ_1 and LPIPS [45] losses:

$$\mathcal{L}_{\text{pixel}} = \|\mathcal{D}(\hat{z}_T) - T\|_1 + \lambda_{\text{lpips}} \cdot \mathcal{L}_{\text{LPIPS}}(\mathcal{D}(\hat{z}_T), T), \tag{6}$$

Inference. At inference, we apply a single forward pass at $t = 0$, so that $\hat{z}_T = z_I + v_\theta(z_I, 0)$ directly. This one-step scheme is consistent with the linear flow matching objective [22], where the ground-truth velocity field is constant along the trajectory; we train and infer at $t = 0$ for simplicity and empirical stability.

3.4 Latent Composition Consistency

While the latent flow matching objective encourages accurate transmission recovery, it does not explicitly prevent entanglement between $\hat{z}_T$ and $\hat{z}_R$: a degenerate solution exists where the model encodes scene-specific information into both components without achieving true disentanglement. We address this with two complementary losses — $\mathcal{L}_{\text{cycle}}$ and $\mathcal{L}_{\text{LCS}}$ — both grounded in the **Latent Composition Consistency (LCC)** strategy described below.

Latent Composition Consistency (LCC). The low mutual coherence of transmission and reflection in latent space (Fig. 1(a)) enables a unique training strategy. Since the two components exhibit substantially lower overlap, composing the transmission of one image with the reflection of another yields a semantically valid mixture — something far less effective in pixel space. A well-disentangled model should therefore decompose not only real mixtures but any such synthetic composition. We operationalize this by cyclically swapping reflection components across the N samples of a training batch (Fig. 2(b)):

$$\tilde{z}_I^i = \hat{z}_T^i + \hat{z}_R^{(i \bmod N)+1}, \quad i \in \{1, \ldots, N\}. \tag{7}$$

A second forward pass on each synthetic mixture then recovers the decomposition:

$$\tilde{v}^i = v_\theta(\tilde{z}_I^i, 0), \qquad \tilde{z}_T^i = \tilde{z}_I^i + \tilde{v}^i, \qquad \tilde{z}_R^i = -\tilde{v}^i. \tag{8}$$

Cycle Consistency Loss ($\mathcal{L}_{\text{cycle}}$). The cycle loss enforces that the second-pass decomposition recovers the original components, averaged over the batch:

$$\mathcal{L}_{\text{cycle}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(\|\tilde{z}_T^i - \texttt{sg}(\hat{z}_T^i)\|_2^2 + \|\tilde{z}_R^i - \texttt{sg}\Big(\hat{z}_R^{(i \bmod N)+1}\Big)\|_2^2\right), \tag{9}$$

where $\texttt{sg}(\cdot)$ is the stop-gradient operator. Without it, the model could trivially minimize $\mathcal{L}_{\text{cycle}}$ by collapsing all first-pass predictions to the same value (mode collapse). Stop-gradient fixes the first-pass outputs as stable targets, so that the second pass learns to match a consistent decomposition anchored by $\mathcal{L}_{\text{latent}}$ and $\mathcal{L}_{\text{pixel}}$.

Layer Contrastive Separation Loss ($\mathcal{L}_{\text{LCS}}$). While $\mathcal{L}_{\text{latent}}$ supervises the velocity target $v^* = z_T - z_I$ (which only requires $\mathcal{E}(T)$ and $\mathcal{E}(I)$), it does not explicitly enforce that the resulting $\hat{z}_T$ and $\hat{z}_R$ are semantically disentangled. Introducing an explicit reflection regression target $\mathcal{E}(I-T)$ would be problematic, since $\mathcal{E}(I-T) \neq \mathcal{E}(I) - \mathcal{E}(T)$ due to encoder nonlinearity. Instead, we exploit the low mutual coherence of z_T and z_R (cosine similarity 0.07 in latent vs. 0.74 in pixel space) and enforce semantic separation via contrastive learning, without requiring any explicit reflection target. As illustrated in Fig. 2(c), we build on the InfoNCE framework [30], using the cycle-recovered transmission $\tilde{z}_T$ as an anchor, with transmission latents ($\hat{z}_T$, z_T^{gt}) as positives and reflection latents ($\hat{z}_R$, $\tilde{z}_R$, z_R^{gt}) as negatives.

Since reflection and transmission exhibit spatially heterogeneous entanglement, we extract **patch-based** representations: the latent map is divided into non-overlapping $p \times p$ patches, each independently pooled and ℓ_2-normalized, yielding $K = (H_z/p) \times (W_z/p)$ contrastive pairs per sample. Letting $f_k(\cdot)$ denote the patch-pool operation for the k-th patch, and $\sigma_k(z) = f_k(\tilde{z}_T)^\top f_k(z)/\tau$ the scaled similarity to the anchor:

$$\mathcal{L}_{\text{LCS}} = -\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\log\frac{\sum_{s\in\mathcal{P}} e^{\sigma_k(z^s)}}{\sum_{s\in\mathcal{P}} e^{\sigma_k(z^s)} + \sum_{n\in\mathcal{N}} e^{\sigma_k(z^n)}}\right], \tag{10}$$

where $\mathcal{P}$ and $\mathcal{N}$ are the positive and negative sets, $\tau = 0.1$ is the temperature, and $H_z = H/8$, $W_z = W/8$. We set $p = 4$ as the default patch size (see ablation in Sec. 4.3). GT reflection negatives are included only when ground-truth annotations are available.

3.5 Training Objective

The final training objective combines all four loss terms:

$$\mathcal{L} = \lambda_{\text{lat}}\mathcal{L}_{\text{latent}} + \lambda_{\text{pix}}\mathcal{L}_{\text{pixel}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}} + \lambda_{\text{LCS}}\mathcal{L}_{\text{LCS}}, \tag{11}$$

where λ_{lat}, λ_{pix}, λ_{cycle}, and λ_{LCS} are weighting coefficients that balance the contribution of each term. The latent and pixel losses provide direct supervision for transmission recovery, while the cycle and Layer Contrastive Separation losses promote disentanglement without requiring pixel-space reflection ground truth.

Table 1: Benchmark results of various SIRR methods on Real(20), Object(200), Postcard(199), Wild(55), Nature(20), and SIR^2(454) datasets.

Methods	Real(20)		Object(200)		Postcard(199)		Wild(55)		Nature(20)		SIR^2(454)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ERRNet (CVPR'19)	22.89	0.803	24.87	0.896	22.04	0.876	24.25	0.853	20.58	0.756	23.41	0.874
IBCLN (CVPR'20)	21.86	0.762	24.87	0.893	23.39	0.875	24.71	0.886	23.57	0.786	24.08	0.875
LASIRR (ICCV'21)	23.34	0.812	24.36	0.898	23.72	0.903	25.73	0.902	23.45	0.808	24.18	0.893
YTMT (NeurIPS'21)	23.26	0.806	24.87	0.896	22.91	0.884	25.48	0.890	23.85	0.810	24.04	0.880
RobustSIRR (CVPR'23)	23.61	0.835	24.90	0.917	19.91	0.868	23.67	0.884	20.97	0.764	22.54	0.884
DSRNet (ICCV'23)	23.91	0.818	26.74	0.920	24.83	0.911	26.11	0.906	25.22	0.832	25.72	0.907
RRW (CVPR'24)	23.82	0.817	26.55	0.927	24.03	0.903	26.51	0.913	25.96	0.843	25.40	0.908
DSIT (NeurIPS'24)	25.22	0.836	27.27	**0.932**	25.58	**0.922**	27.40	0.918	26.77	0.847	26.50	**0.919**
RDNet (CVPR'25)	25.58	0.846	26.78	0.921	26.33	**0.922**	27.70	0.915	26.21	0.842	26.63	0.915
DAI (AAAI'26)	25.24	0.840	27.26	0.920	27.30	**0.922**	27.44	0.912	27.05	0.846	27.30	**0.919**
PRISM (Ours)	**27.14**	**0.853**	**27.61**	0.925	**27.85**	0.906	**29.54**	**0.932**	**27.35**	**0.853**	**27.95**	0.918

4 Experiment

4.1 Experimental Setup

Implementation Details. Our model is implemented in PyTorch and trained on a single NVIDIA A6000 GPU with a batch size of 2 and a crop size of 512×512 for 50,000 iterations. We use AdamW [24] with cosine annealing [25], decaying from 5×10^{-5} to 1×10^{-6}, in bfloat16 precision. We adopt FLUX.2 Klein 4B [18] as the backbone. Since our task does not involve text conditioning, we replace the text cross-attention input with a single zero-valued embedding token, effectively operating the transformer unconditionally; the text encoder is discarded to save memory. Loss weights are $\lambda_{\text{lat}} = 1$, $\lambda_{\text{pix}} = 1$, $\lambda_{\text{lpips}} = 2$, $\lambda_{\text{cycle}} = 1$, $\lambda_{\text{LCS}} = 0.1$. During training, paired images are augmented with random resize, horizontal flip, rotation, and a random 512×512 crop, applied consistently to input and ground truth.

Training Data. We follow the same training data setting and synthesis pipeline as RDNet [47]: 7,643 synthetic pairs from PASCAL VOC [7], 90 real pairs from [46], and 200 real pairs from Nature [19], sampled with a $0.6 : 0.2 : 0.2$ ratio per epoch. Real pairs provide additional diversity in reflection type and scene complexity not captured by the synthetic distribution.

Evaluation Datasets. We evaluate PRISM on six standard benchmark datasets following the evaluation protocol of RDNet [47]. **Real** [46] consists of 20 challenging real-world reflection images captured in diverse conditions. **SIR^2** [35] is a benchmark comprising three subsets: Object (200 images), Postcard (199 images), and Wild (55 images), covering a wide range of scenes, illumination conditions, and glass types. **Nature** [5] contains 20 real-world samples captured under natural lighting conditions. For in-the-wild evaluation, we additionally use **OpenRR 1K** [44], a real-world reflection removal benchmark; we evaluate on its test set of 100 images in an inference-only setting without any fine-tuning.

Evaluation Metrics. We report PSNR and SSIM [38] on all benchmarks, following prior works [13, 47], computed on the predicted transmission against ground truth in $[0, 1]$. For OpenRR 1K, we additionally report LPIPS [45] and DISTS [4], which measure perceptual similarity and correlate more strongly with

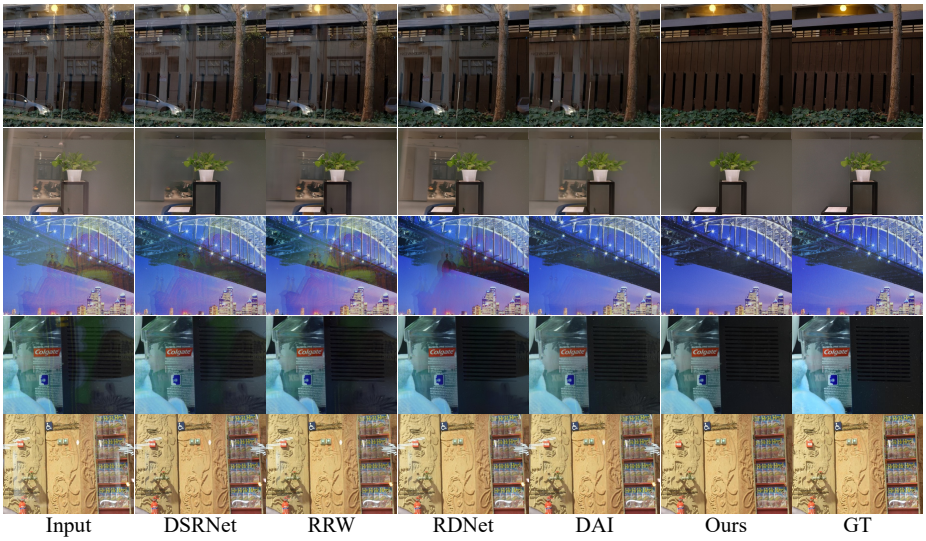

Fig. 3: Visual comparison of different methods on multiple datasets. From top to bottom: Real, Nature, Postcard, Object, and Wild datasets. Zoom in for better visualization of detailed differences.

human visual judgment, as well as NIQE [28] as a no-reference image quality metric.

4.2 Comparison with State-of-the-Art Methods

We compare PRISM against ten state-of-the-art SIRR approaches: ERRNet [39], IBCLN [19], LASIRR [5], YTMT [12], RobustSIRR [16], DSRNet [13], RRW [50], DSIT [14], RDNet [47], and DAI [11]. In all tables, **bold** and underline denote the best and second-best results, respectively.

Quantitative Comparisons. As shown in Tab. 1, PRISM achieves state-of-the-art PSNR across all six benchmarks. PRISM surpasses the previous best RDNet [47] by **1.56 dB** on Real and **1.84 dB** on Wild, with consistent gains on all remaining benchmarks. PRISM also outperforms the diffusion-based DAI [11] on all reported datasets, demonstrating that structured latent decomposition provides a strong inductive bias for reflection removal that generalizes well across diverse real-world conditions. On SSIM, PRISM achieves the best scores on Real, Wild, and Nature, while trailing DSIT [14] and DAI [11] slightly on Object and Postcard. This pattern is consistent with the known perception-distortion tradeoff [1]: generative priors tend to recover sharper, higher-fidelity outputs that improve average pixel accuracy (PSNR) but may introduce local high-frequency details that deviate from the ground truth's exact spatial structure, which SSIM's window-based structural term penalizes. This observation is further corroborated

Fig. 4: Qualitative comparison of estimated reflection layers. The top row shows a solid-object scene and the bottom row shows a wild scene, both from the SIR^2 dataset [35]. Best viewed zoomed in.

in Tab. 2, where PRISM achieves substantially better perceptual scores (LPIPS -14.0%, DISTS -14.6% relative to RDNet) despite slightly lower SSIM — metrics known to correlate more strongly with human visual judgment [4, 45].

Qualitative Comparisons. Fig. 3 shows visual results against DSRNet [13], RRW [50], RDNet [47], and DAI [11] on five benchmarks. Across all datasets, competing methods tend to either leave residual reflection artifacts or introduce color distortion and structural degradation in the recovered transmission. Notably, in the last row, residual reflection patterns remain visible even in the ground-truth image, whereas PRISM successfully removes these artifacts, yielding a cleaner and more consistent result, further demonstrating the advantage of our latent-space decomposition over pixel-space approaches. Additional visual results are provided in the supplementary material.

Fig. 4 compares the estimated reflection layers of YTMT [12], DSRNet [13], DSIT [14], and RDNet [47] against ours. All competing methods produce near-zero reflection estimates, indicating that pixel-space approaches tend to absorb the reflection component into reconstruction residuals rather than explicitly separating it. This is a fundamental limitation of pixel-space decomposition: without a structured additive prior, the reflection branch degenerates into a trivial zero output while all information is passed to the transmission branch. PRISM, by contrast, recovers structurally coherent reflections that preserve the spatial layout of the ground truth. Although the recovered reflections exhibit brightness differences due to the absence of pixel-level reflection supervision, the clear structural correspondence demonstrates that latent-space decomposition achieves more effective layer disentanglement than pixel-space alternatives, even without a dedicated pixel-space reflection loss.

In-the-wild Comparison. We further evaluate the generalization capability of PRISM on the in-the-wild OpenRR 1K dataset [44]. As shown in Tab. 2, all models are evaluated on the OpenRR 1K test set without being trained on this dataset, *i.e.*, in an inference-only setting. PRISM achieves the best performance in PSNR, LPIPS, DISTS, and NIQE among all compared methods.

Table 2: Quantitative comparison on the OpenRR 1K test set [44]. All methods are evaluated in an inference-only setting without training on this dataset. Inference is performed on 512 × 512 images, and inference time is measured on an NVIDIA RTX 4090 GPU.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	Time (ms)↓	Params
YTMT	25.16	0.927	0.096	0.059	3.6241	<u>75.68</u>	<u>58.47M</u>
DSRNet	26.24	<u>0.935</u>	<u>0.074</u>	<u>0.048</u>	3.6573	288.01	144.65M
DSIT	26.01	0.903	0.083	0.051	<u>3.5536</u>	222.91	326.96M
RRW	25.55	0.929	0.086	0.055	3.6825	**35.88**	**27.99M**
RDNet	<u>26.81</u>	**0.939**	<u>0.074</u>	<u>0.048</u>	3.7713	102.26	314.03M
DAI	26.04	0.903	0.085	0.070	3.6197	138.70	1.30B
PRISM	**27.23**	0.893	**0.064**	**0.041**	**3.5512**	104.23	3.97B

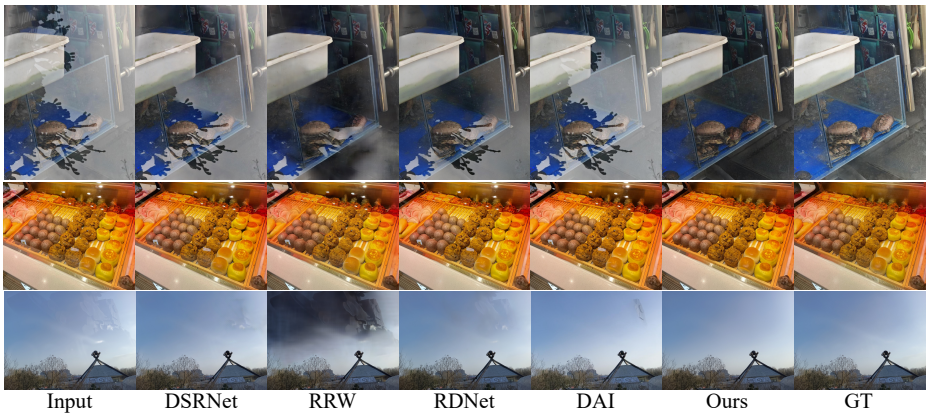

Fig. 5: Visual comparison on the OpenRR 1K test set [44]. All methods are evaluated without training on this dataset.

While SSIM is lower than RDNet, PRISM outperforms all baselines on perceptual metrics (LPIPS −14.0%, DISTS −14.6% relative to RDNet), consistent with the perception-distortion tradeoff discussed above.

Tab. 2 also reports the number of parameters and inference time. Despite having more parameters, PRISM achieves comparable inference time to RDNet (104 ms vs. 102 ms), as `bf16` precision substantially reduces memory bandwidth and computation overhead.

Fig. 5 provides visual comparisons. PRISM effectively captures reflection patterns and removes them while preserving the underlying transmission structures. Compared with competing methods, PRISM produces cleaner results with fewer residual artifacts, demonstrating stronger generalization in real-world reflection removal scenarios.

Table 3: Ablation study of $\mathcal{L}_{\text{cycle}}$ and $\mathcal{L}_{\text{LCS}}$. The base model (row 1) uses the velocity target $v^* = z_T - z_I$ supervised by $\mathcal{L}_{\text{latent}}$ and $\mathcal{L}_{\text{pixel}}$ only, without explicit disentanglement losses.

$\mathcal{L}_{\text{cycle}}$	$\mathcal{L}_{\text{LCS}}$	**Real(20)**		**Nature(20)**		**SIR^2(454)**		**Avg**	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
×	×	26.67	0.847	27.24	0.850	27.68	0.916	27.20	0.871
✓	×	26.90	**0.853**	**27.35**	0.851	27.86	0.917	27.37	0.874
×	✓	26.74	0.846	27.34	0.852	27.90	0.917	27.33	0.872
✓	✓	**27.14**	**0.853**	**27.35**	**0.853**	**27.95**	**0.918**	**27.48**	**0.875**

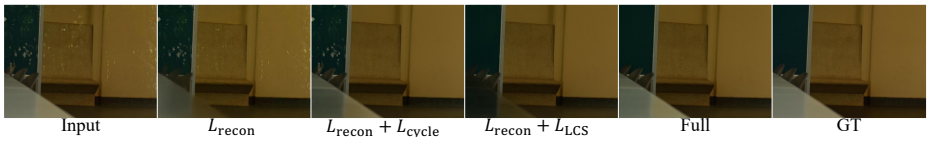

Fig. 6: Progressive ablation of each loss component. $\mathcal{L}_{\text{recon}} := \mathcal{L}_{\text{latent}} + \mathcal{L}_{\text{pixel}}$. Each column adds one loss on top of the previous setting.

4.3 Ablation Study

We conduct ablation studies on Real(20), Nature(20), and SIR^2(454) to analyze each proposed component.

Effect of $\mathcal{L}_{\text{cycle}}$ and $\mathcal{L}_{\text{LCS}}$. Tab. 3 isolates the contribution of each disentanglement objective. The base model (row 1) uses only $\mathcal{L}_{\text{latent}}$ and $\mathcal{L}_{\text{pixel}}$, confirming that the additive latent formulation provides a useful inductive bias on its own; however, without explicit disentanglement losses, the model may still entangle scene-specific information across the two components. Adding $\mathcal{L}_{\text{cycle}}$ yields consistent improvements across all three benchmarks (+0.23 dB on Real, +0.18 dB on SIR^2), confirming that enforcing reconstruction consistency under cyclically swapped reflections strengthens disentanglement. Combining both losses further improves performance, with the full model achieving 27.14 / 27.35 / 27.95 dB on Real, Nature, and SIR^2 respectively, demonstrating that $\mathcal{L}_{\text{cycle}}$ and $\mathcal{L}_{\text{LCS}}$ are complementary: the cycle loss enforces structural consistency in latent space, while $\mathcal{L}_{\text{LCS}}$ provides a semantic disentanglement signal without relying on exact additive reconstruction. Fig. 6 visualizes the progressive effect: $\mathcal{L}_{\text{recon}}$ alone leaves residual reflections; adding $\mathcal{L}_{\text{cycle}}$ stabilizes the transmission but reflections persist; $\mathcal{L}_{\text{LCS}}$ suppresses them further yet introduces minor artifacts; only the full combination removes reflections while preserving fidelity.

Effect of Cycle Supervision Target. Tab. 4 investigates which components are supervised within the cycle branch. Supervising only the reflection cycle ($\tilde{z}_R$) or only the transmission cycle ($\tilde{z}_T$) in isolation both degrade performance

Table 4: Ablation study on the cycle supervision target. $\tilde{z}_T$ and $\tilde{z}_R$ indicate whether the transmission and reflection components of the cycle branch are supervised, respectively.

$\tilde{z}_T$	$\tilde{z}_R$	**Real(20)**		**Nature(20)**		**SIR^2(454)**		**Avg**	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
×	✓	<u>26.95</u>	<u>0.850</u>	27.26	0.851	<u>27.87</u>	<u>0.917</u>	<u>27.36</u>	<u>0.873</u>
✓	×	26.86	0.849	<u>27.32</u>	<u>0.852</u>	27.83	<u>0.917</u>	27.34	0.872
✓	✓	**27.14**	**0.853**	**27.35**	**0.853**	**27.95**	**0.918**	**27.48**	**0.875**

Table 5: Ablation study on spatial representation strategy for the Layer Contrastive Separation loss. Patch (p) denotes non-overlapping patch-average-pool with patch size p; Global Avg. Pool collapses the entire spatial map to a single vector.

Representation	**Real(20)**		**Nature(20)**		**SIR^2(454)**		**Avg**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Patch ($p = 2$)	<u>27.03</u>	<u>0.851</u>	**27.35**	**0.853**	<u>27.94</u>	<u>0.917</u>	<u>27.44</u>	<u>0.874</u>
Patch ($p = 4$, PRISM)	**27.14**	**0.853**	**27.35**	**0.853**	**27.95**	**0.918**	**27.48**	**0.875**
Patch ($p = 8$)	26.83	0.849	<u>27.22</u>	<u>0.852</u>	27.93	<u>0.917</u>	27.33	0.873
Global Avg. Pool	26.73	0.845	27.21	<u>0.852</u>	<u>27.94</u>	**0.918**	27.29	0.872

compared to supervising both jointly, underscoring the importance of bidirectional consistency. Jointly supervising both components achieves the best results across all benchmarks, as each direction provides a complementary constraint: the transmission cycle enforces that the scene content is preserved under arbitrary reflection corruption, while the reflection cycle ensures that the separated reflection latent is consistent with the physical composition.

Effect of Patch-Based Representation. Tab. 5 compares spatial representation strategies for $\mathcal{L}_{\text{LCS}}$. Global average pooling performs worst (26.73 dB on Real), as collapsing the entire latent map into a single vector discards the spatially heterogeneous entanglement structure. Patch-based representations consistently outperform global pooling across all tested sizes, with $p = 4$ achieving the best balance between spatial granularity and representation stability (27.14 / 27.95 dB on Real / SIR^2). Too-small patches ($p=2$) introduce noise from overly local comparisons, while too-large patches ($p = 8$) lose the fine-grained spatial cues needed to distinguish locally entangled layers.

5 Conclusion

In this paper, we presented PRISM, a latent flow matching framework for single-image reflection removal that reinterprets the task as a latent linear separation problem. By exploiting the low mutual coherence of pretrained VAE latent representations, PRISM adopts an approximate additive formulation and recovers

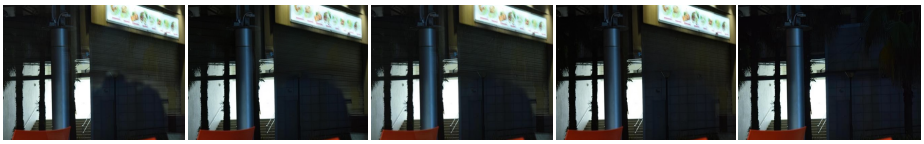

Fig. 7: Failure case: dense and saturated reflection on SIR^2 [35]. From left to right: input I, RDNet [47], DAI [11], Ours, and ground truth T.

the transmission via a velocity field in a single forward pass, without dual-branch architectures or a dedicated pixel-space reflection loss. To enforce robust disentanglement, PRISM introduces a Latent Composition Consistency strategy — uniquely enabled by the broad decorrelation of latent representations — and a Layer Contrastive Separation loss that handles the non-exact nature of the additive assumption without propagating approximation errors. Extensive experiments on six real-world benchmarks demonstrate that PRISM consistently outperforms state-of-the-art approaches, achieving significant PSNR gains while maintaining strong generalization to challenging real-world imagery.

Limitations. The FLUX backbone (3.97B parameters) incurs significant memory overhead, limiting deployment on resource-constrained hardware. Additionally, while the additive latent formulation is empirically effective, a principled theoretical understanding of when and why the VAE encoder approximately preserves additive structure remains an open question. Fig. 7 illustrates a representative failure mode: when the reflection is optically dense and saturated (*e.g.*, $\|R\|_1/\|I\|_1 > 0.5$), z_I shifts toward the reflection subspace and the low-coherence assumption weakens, causing all methods—including PRISM—to leave residual reflections. We provide further analysis in the supplementary material.

Acknowledgments

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026 (Project Name: Development of AI-Based Animation Production Technology to Ensure Character and Scene Consistency and Continuity for Enhanced Efficiency and Quality, Project Number: RS-2026-25525207, Contribution Rate: 30%). This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00156, Fundamental research on continual meta-learning for quality enhancement of casual videos and their 3D metaverse transformation), IITP grant funded by the Korea government (MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University)).

References

1. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR (2018)
2. Chang, Y., Jung, C., Sun, J.: Joint reflection removal and depth estimation from a single image. IEEE Transactions on Cybernetics (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
4. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE TPAMI (2020)
5. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: ICCV (2021)
6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
8. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: ICCV (2017)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
11. Hu, J., Yang, C., Zhou, Z., Fang, J., Yang, X., Tian, Q., Shen, W.: Dereflection any image with diffusion priors and diversified data. In: AAAI (2026)
12. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation. NeurIPS (2021)
13. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: CVPR (2023)
14. Hu, Q., Wang, H., Guo, X.: Single image reflection separation via dual-stream interactive transformers. NeurIPS (2024)
15. Kee, E., Pikielny, A., Blackburn-Matzen, K., Levoy, M.: Removing reflections from raw photos. In: CVPR (2025)
16. Kim, S., Huo, Y., Yoon, S.E.: Single image reflection removal with physically-based training images. In: CVPR (2020)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Labs, B.F.: Flux 2 klein 4b. `https://blackforestlabs.ai` (2025)
19. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: CVPR (2020)
20. Li, Y., Brown, M.S.: Single image layer separation using relative smoothness. In: CVPR (2014)
21. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: ECCV (2024)
22. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
23. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

25. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: ICLR (2017)
26. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. In: ICML (2023)
27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
28. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Sign. Process. Letters (2012)
29. Nayar, S.K., Fang, X.S., Boult, T.: Separation of reflection components using color and polarization. IJCV (1997)
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
31. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV (2020)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
33. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: CVPR (2015)
34. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML (2023)
35. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: ICCV (2017)
36. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Face image reflection removal. IJCV (2021)
37. Wang, L., Cheng, K., Lei, S., Wang, S., Yin, W., Lei, C., Long, X., Lu, C.T.: Dc-gaussian: Improving 3d gaussian splatting for reflective dash cam videos. NeurIPS (2024)
38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
39. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: CVPR (2019)
40. Wu, C., Li, J., Zhou, J., Lin, J., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
41. Wu, G., Jiang, J., Jiang, K., Liu, X.: Learning from history: Task-agnostic model contrastive learning for image restoration. In: AAAI (2024)
42. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: CVPR (2021)
43. 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: ICCV (2023)
44. 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. In: ICIP. IEEE (2025)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
46. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: CVPR (2018)
47. Zhao, H., Li, M., Hu, Q., Guo, X.: Reversible decoupling network for single image reflection removal. In: CVPR (2025)

48. Zheng, Q., Shi, B., Chen, J., Jiang, X., Duan, L.Y., Kot, A.C.: Single image reflection removal with absorption effect. In: CVPR (2021)
49. Zhu, Y., Zhao, W., Li, A., Tang, Y., Zhou, J., Lu, J.: Flowie: Efficient image enhancement via rectified flow. In: CVPR (2024)
50. 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: CVPR (2024)