340 FPS Reflection-free Video from Spikes Modulated by a Rapidly Rotating Polarizer

Lishuai Huang[1], Yeliduosi Xiaokaiti[2], Youwei Lyu[3], Langyue Chang[1], Boxin Shi[2], Shikui Wei[1], Yao Zhao[1], Meng Jian[4], Yashen Wang[5], and Yakun Chang[1*]

[1]Beijing Jiaotong University, [2]Peking University, [3]vivo,
[4]Beijing University of Technology, [5]CETC
lishuaihuang215@gmail.com, yongqiye@stu.pku.edu.cn, shiboxin@pku.edu.cn, youweilv@gmail.com, {ykchang, shkwei, yzhao}@bjtu.edu.cn

Abstract. Removing reflections from mixed light using polarization offers the benefits of robust physical constraints. However, extending this to high-speed scenes presents challenges due to the discrete frames of conventional cameras. In this paper, we propose 340 FPS reflection-free video reconstruction by leveraging the high-speed spike camera modulated by a rapidly rotating polarizer. The core motivation of our approach relies on exploiting the continuous and asynchronous firing of the spike sensor to densely capture instantaneous polarization states. However, the inherently low signal-to-noise ratio of spikes under extremely short integration windows and light attenuation causes severe numerical instability in physical layer separation. To address this, we introduce a physically-driven framework featuring a differentiable Reliability-Weighted Physics Solver (RWPS). RWPS robustly separates reflection and transmission layers by dynamically adjusting learned confidence weights to handle severe shot noise. Furthermore, we develop a refinement module that leverages back-projection residuals and physical-discrepancy cues to accurately restore high-frequency textures. Extensive experiments demonstrate that our method enables robust reconstruction of high-fidelity and reflection-free videos.

Keywords: Reflection removal · spike camera · polarization · high-speed imaging · video reconstruction.

*Corresponding author.
[1]Lishuai Huang, Langyue Chang, Shikui Wei, Yao Zhao, and Yakun Chang are with Institute of Information Science, Beijing Jiaotong University and Visual Intelligence +X International Cooperation Joint Laboratory of the MoE. [2]Yeliduosi Xiaokaiti and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University. Boxin Shi is also with PKU-AI[2] Robotics Joint Lab of Embodied AI. [3]Youwei Lyu is with vivo BlueImage Lab, vivo Mobile Communication Co., Ltd. [4]Meng Jian is with School of Information Science and Technology, Beijing University of Technology. [5]Yashen Wang is with Artificial Intelligence Institute of CETC.

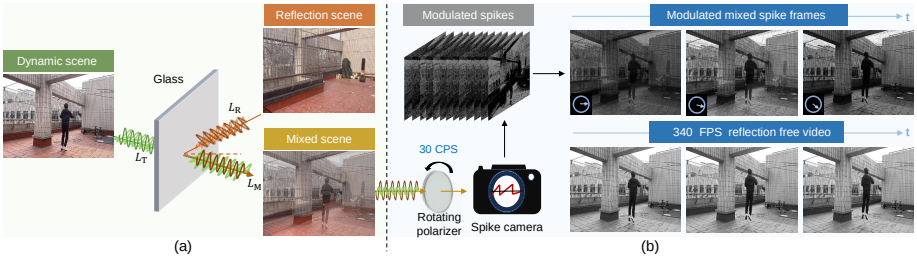

Fig. 1: High frame rate reflection-free video reconstruction via a rapidly rotating polarizer. (a) A spike camera coupled with a rapidly rotating polarizer continuously captures the dynamic polarization states of the mixed scene (L_{M} is a mixture of transmission L_{T} and reflection L_{R}) as ultra-high-speed spikes. (b) By decoupling the modulated spikes, we reconstruct 340 FPS reflection-free videos.

1 Introduction

When capturing images or videos through glass, the incident light L_{M} is a mixture of the reflection light L_{R} and the transmission light L_{T}. Recovering L_{T} from L_{M} is an ill-posed problem, as there are an infinite number of solutions in the absence of any prior knowledge. Pioneering works address this ambiguity by leveraging statistical priors such as gradient sparsity [15], relative smoothness [17], and ghosting artifacts [26]. Following the success of deep learning, data-driven methods [2, 4, 5, 9, 28, 30, 37] reduce the reliance on manually designed statistical priors. More recently, the large language model guided reflection removal [6, 40] demonstrates the capability for transmission synthesis under strong reflections.

However, generative methods often rely on estimating plausible textures rather than ensuring physical consistency with the original scene. To recover the actual transmission layer, researchers use physical cues like flash [3, 7, 13] and polarization [14, 20–22] to obtain observable fluctuations in the reflection or transmission layers. Unlike active flash that only works within several meters [3, 13], passive polarization-based methods offer the flexibility regarding the effective imaging distance. This advantage makes them more effective for reflection removal in the wild. However, it is often difficult to capture the exact state corresponding to Brewster's angle, *i.e.*, the specific orientation where the reflection is at its minimum. Therefore, existing solutions rely on several discrete angles [11, 12] to construct a linear system, or leverage deep learning to remove the reflections from a pair of polarized and unpolarized images [20, 21].

According to the properties of overdetermined linear systems, if we can continuously capture every state of the polarizer during its rotation, we could construct a robust and easily solvable linear system. However, this is difficult to achieve with conventional frame-based cameras due to the discrete nature of frames. Although a micro-polarizer camera such as PolarCam [1] can simultaneously capture four polarization states with a single shot, it is constrained by a limited frame rate of 24 FPS. Unlike conventional cameras, the new generation

of high-speed spike cameras [10] can continuously sense environments at a rate of more than 20,000 Hz. This allows to capture every instantaneous light polarization state when rotating a polarizer. As shown in Fig. 1, through our mechanical design, the polarizer can rotate at a high speed of 30 cycles per second (CPS). Consequently, the rapid rotation of a polarizer combined with spike sensing makes the reconstruction of high-frame-rate and reflection-free video achievable. However, since spikes are discrete and binary representations of incident photons, directly applying continuous polarization physics to this asynchronous modality remains an open problem. Second, the light attenuation induced by a polarizer degrades the signal-to-noise ratio (SNR). Therefore, how to robustly reconstruct reflection-free videos from the noisy and modulated spikes is highly non-trivial.

To address these challenges, we propose a physically-driven framework designed for reflection removal under rapid polarization modulation. We first formulate a specialized spike imaging model for the rotating polarizer setup, bridging the gap between the binary spike representation and the continuous incident mixed light. To handle the severe noise, we design a differentiable Reliability-Weighted Physics Solver (RWPS). RWPS separates the reflection and transmission layers by dynamically adjusting a set of learned weights derived from a lightweight network, explicitly conditioned on the reliability of the observed spike modulation. Finally, because the baseline results suffer from residual noise and oversmoothed details, we introduce a refinement module. This module leverages back-projected physical residuals to enhance the overall visual fidelity. Extensive experimental results on both synthetic and real-scene data demonstrate the capability of reconstructing reflection-free videos. The main contributions of this paper are summarized as follows:

- a high-speed imaging system that combines a spike camera with a rapidly rotating polarizer to capture dense temporal polarization states, enabling the reconstruction of 340 FPS reflection-free video;
- a differentiable RWPS that robustly removes reflections by dynamically adjusting learned confidence weights to handle noise in high-speed spikes; and
- a refinement module that leverages back-projection residuals and physical-discrepancy cues to restore high-frequency textures, supported by a new dataset specifically designed for our high-speed setup.

2 Related work

Image-based methods. Image-based methods initially tackled this by imposing hand-crafted priors, such as gradient sparsity [15] and ghosting cues [26]. Deep learning methods, such as CRRN [28], integrate multi-scale gradient information to guide the separation. To handle complex blending, IBCLN [16] employs recurrent refinement, while RAGNet [19] utilizes reflection-aware guidance in a two-stage framework. Architectural innovations have further pushed the boundaries: RDNet [38] leverages a reversible architecture to preserve information flow, and Hu *et al.* [9] exploit the synergy between transmission and

reflection components to resolve residual ambiguities. Song *et al*. [27] incorporate cross-scale attention and adversarial discrimination to defend against adversarial perturbations. Some methods also revisit physics-inspired priors within a single-image context, such as explicitly modeling the absorption effect [39] to approximate refractive amplitude. Transformers have been adopted to capture long-range context [18], and diffusion models are employed to hallucinate high-frequency transmission details [35]. To mitigate semantic ambiguity where reflections resemble valid objects, Language-guided methods [6, 40] introduce text embeddings as auxiliary supervision. When multiple frames are available, methods like [23, 32] exploit motion information to distinguish the background from the moving reflection layer. Despite these advances, image-based methods inherently struggle when the reflection is strong and sharp. Without definitive physical evidence (such as polarization state or active lighting modulation), pure image-based solutions often fail to distinguish valid object textures.

Physically driven methods. Physically driven methods leverage extra hardware cues to resolve the inherent ambiguity of layer separation. Early works formulate linear constraints based on the Fresnel equations to separate layers [25]. Deep learning approaches have since incorporated these cues: Lyu *et al*. [20] combine unpolarized-polarized pairs with a physics-based refinement network, while Lei *et al*. [14] introduce a robust pipeline to handle misalignment in handheld polarized capture. Most recently, PolarFree [34] combines polarization cues with diffusion priors to enable reflection-free imaging in dynamic scenes. Chang *et al*. [3] proposed a Siamese network to process flash/no-flash pairs, while Lei *et al*. [13] extended this to more robust reflection removal using flash-only cues. Depth and Focus cues leverage the focal disparity between the reflection and the transmission layer. Wan *et al*. [29] utilize depth-of-field cues to label edges, while Punnappurath *et al*. [24] exploit dual-pixel sensor data, treating the slight defocus disparity between left and right photodiodes as a stereo cue to separate layers based on depth. However, a limitation across these approaches is their reliance on discrete frame acquisition, which restricts their applicability to static or slow-moving subjects. In dynamic scenarios, the motion occurring between sequential exposures breaks the physical consistency required for alignment. Therefore, in this paper, we fill this gap by introducing a high-speed spike vision system capable of capturing continuous polarization modulation, thereby extending reflection removal capabilities to more general dynamic scenes.

3 Method

We propose a physically-driven framework for reconstructing high-frame-rate, reflection-free video from spikes. In Sec. 3.1, we introduce the polarized spike imaging model and the relative zero-motion observation. In Sec. 3.2 and Sec. 3.3, we detail our two-stage framework, consisting of a baseline removal stage to reconstruct initial video and a refinement module to faithfully restore high-

frequency textures. Finally, in Sec. 3.4 and Sec. 3.5, we introduce loss functions and real-scene data.

3.1 Preliminaries

Polarized spike imaging model. Let $L_{\mathrm{T}}(p,t)$ and $L_{\mathrm{R}}(p,t)$ denote the transmission and reflection light intensities at pixel p and timestamp t. As the incident light passes through a continuously rotating polarizer, its intensity is dynamically modulated. By Fresnel equations, the reflected light is partially linearly polarized, while the transmitted background light is mostly unpolarized. The mixed light intensity $L_{\mathrm{M}}(p,t)$ is formulated as:

$$L_{\mathrm{M}}(p,t) = \frac{1}{2}\left(\zeta(p,t)L_{\mathrm{R}}(p,t) + (1-\zeta(p,t))L_{\mathrm{T}}(p,t)\right), \tag{1}$$

where the factor $1/2$ accounts for the linear polarizer's intrinsic attenuation of unpolarized light. The time-varying mixing coefficient $\zeta(p,t) \in [0,1]$ modulates the reflection strength via Fresnel reflectance components $R_{\perp}$ and $R_{\parallel}$ [11]:

$$\zeta(p,t) = R_{\perp}(p,t)\cos^2(\phi(p,t) - \phi_{\perp}(p,t)) + R_{\parallel}(p,t)\sin^2(\phi(p,t) - \phi_{\perp}(p,t)), \tag{2}$$

where $\phi(p,t)$ denotes the instantaneous angle of the rotating polarizer and $\phi_{\perp}(p,t)$ denotes the projected incidence angle. Leveraging the ultra-high temporal resolution (*e.g.*, 20,000 Hz), spike cameras are capable of capturing $L_{\mathrm{M}}(p,t)$ while tracking the high-speed dynamics of $\phi(p,t)$. Each pixel continuously integrates the instantaneous modulated light $L_{\mathrm{M}}(p,\tau)$: $A(p,t) = \int_{t_{\mathrm{prev}}}^{t} \alpha L_{\mathrm{M}}(p,\tau)d\tau$, where $A(p,t)$ is the accumulated photon-generated electrons, α is the photosensitivity, and t_{prev} is the previous spike timestamp. When $A(p,t)$ reaches a predefined threshold, a spike $S(p,t) = 1$ is fired and the accumulator resets.

Relative zero-motion observation. Since spike cameras utilize a 1-bit data representation rather than a traditional image format, we first reconstruct intensity maps by accumulating spikes over a fixed temporal window Δt to ensure compatibility with mainstream deep learning frameworks. This reconstruction is

$$\mathbf{I}(p,t) = \frac{1}{\Delta t}\sum_{\tau=t-\Delta t/2}^{t+\Delta t/2} S(p,\tau). \tag{3}$$

For brevity, we denote $\mathbf{I}(p,t)$ as $\mathbf{I}(i)$, omitting the spatial coordinate p and utilizing i as the sequence index to replace t. Similarly, $\phi(p,t)$ is simplified to $\phi(i)$. Correspondingly, let $\mathbf{R}(i)$ and $\mathbf{T}(i)$ denote the digital image representations of the latent light sources $L_{\mathrm{R}}(p,t)$ and $L_{\mathrm{T}}(p,t)$. In our default setting, $\Delta t = 0.65$ ms, adjacent reconstruction centers are sampled with an angular stride of $31.5°$, so the output cadence is $(30 \times 360)/31.5 \approx 343$ FPS, which we report as 340 FPS. Given a sequence of $2K+1$ observations $\{\mathbf{I}(i+k), \phi(i+k) | k \in [-K, K]\}$, we introduce the relative zero-motion observation: Within a small temporal window and a rapidly rotating polarizer, the object motions are negligible, while the

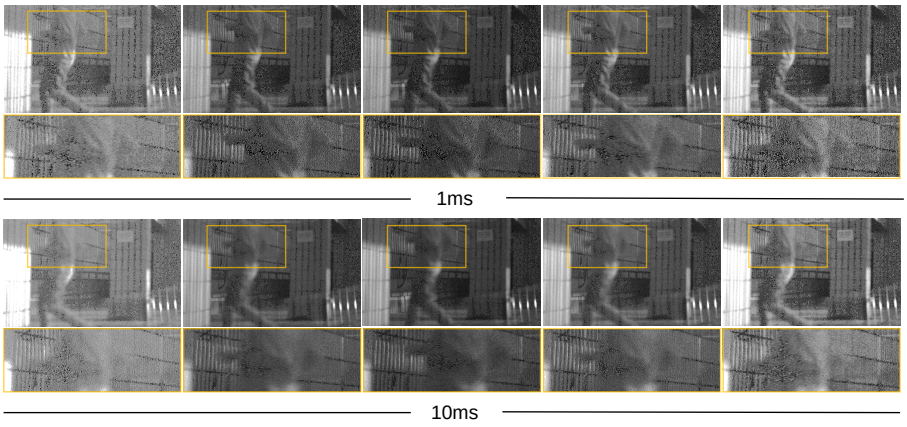

Fig. 2: Relative zero-motion observation. At a highly restricted temporal window of $\Delta t = 1$ ms, motion blur across consecutive spike frames is negligible, while the intensity attenuation from the rotating polarizer remains highly observable. A longer observation window (10 ms) incurs more motion blur.

modulation of polarized light intensity remains observable. As shown in Fig. 2, this observation provides the essential baseline for the construction of a linear system. Specifically, it enables the removal of the reflection layer $\mathbf{R}(i)$ from the $2K+1$ frames. This observation is relative rather than absolute. Our supplementary analysis further shows that the default observation window remains stable under moderate shifts, but degrades when the accumulated shift becomes large[2].

3.2 Baseline removal

As illustrated in Fig. 4, our framework employs a two-stage architecture consisting of baseline removal and subsequent refinement. In this section, we first detail the baseline removal stage. The relative zero-motion observation suggests that reconstructing a clean transmission layer $\mathbf{T}(i)$ is theoretically feasible by solving a system of linear equations. To achieve this, we first rewrite the imaging model into a linear equation:

$$2\mathbf{I}(i) + \epsilon(i) = \zeta(i)\mathbf{R}(i) + (1 - \zeta(i))\mathbf{T}(i), \tag{4}$$

where $\epsilon(i)$ denotes the noise term. Before stacking equations for all $2K+1$ frames, to account for potential micro-motions within the sequence, we spatially align the adjacent frames to the center frame i using a Reflection-Invariant Registration (RIR)[3] module [33]. RIR first fits a harmonic response over the local stack and estimates a Degree of Polarization (DoP) map. Since reflection-dominant pixels usually have stronger polarization modulation, RIR masks out high-DoP regions

[2] More experimental analysis is provided in the supplementary material.
[3] More technical details are provided in the supplementary material.

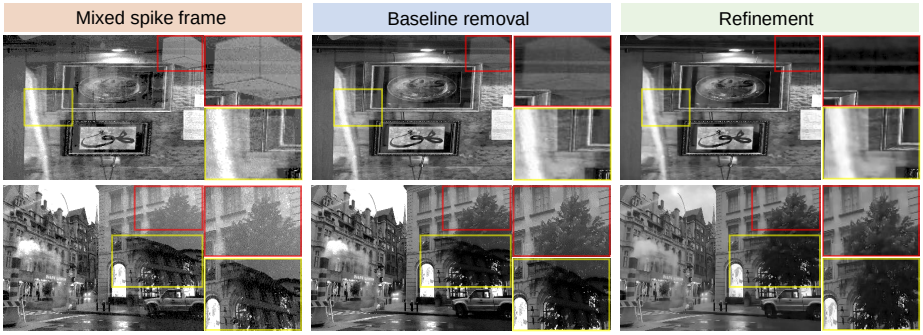

Fig. 3: The boxed regions show that reflections are not completely removed in the baseline results, and some high-frequency details become blurred. We employ the refinement module discussed in Sec. 3.3 to remove the residual reflections and achieve better detail reconstruction.

and performs masked phase correlation on the more stable transmission background. After alignment, we obtain the matrix form $\mathbf{Ax} \approx \mathbf{b}$, where $\mathbf{x} = [\mathbf{R}, \mathbf{T}]^\top$ is the target vector, $\mathbf{b} = [2\mathbf{I}(i-K), \ldots, 2\mathbf{I}(i+K)]^\top$ is the observation vector, and $\mathbf{A} \in \mathbb{R}^{(2K+1)\times 2}$ is the coefficient matrix derived from $\{\zeta(i-K), \ldots, \zeta(i+K)\}$. However, high-speed imaging presents two significant technical challenges. First, the heavy noise caused by the narrow temporal window will significantly degrade the signal-to-noise ratio of the results. Second, within the localized temporal sequence $\{\phi(i+k) \mid k \in [-K, K]\}$, the variation in the polarization angle is relatively marginal. To address these challenges, we propose a differentiable Reliability-Weighted Physics Solver (RWPS).

Reflection removal with RWPS. Not all observations contribute equally to a robust solution. To suppress outliers and noise, we introduce a diagonal weight matrix $\mathbf{W} = \mathrm{diag}(w(i-K), \ldots, w(i+K))$ where each diagonal element $w(\cdot)$ denotes the reliability weight of the corresponding polarization-modulated spike observation, with values constrained to $[0, 1]$ to quantify the confidence level of each frame in the linear system solving. Both this weight matrix $\mathbf{W}$ and the polarization mixing coefficients constructing the matrix $\mathbf{A}$ are jointly predicted by a lightweight parameter estimation network (Sec. 3.3). We minimize the reliability-weighted ridge regression objective:

$$\mathcal{L}(\mathbf{x}) = ||\mathbf{W}^{1/2}(\mathbf{Ax} - \mathbf{b})||_2^2 + \lambda||\mathbf{x}||_2^2, \tag{5}$$

where λ is set per pixel from the **Conf** (Sec. 3.3) as $\lambda = \lambda_{\text{low}}\mathbf{Conf} + \lambda_{\text{high}}(1 - \mathbf{Conf})$. Thus, unreliable pixels receive stronger regularization. The closed-form solution $\hat{\mathbf{x}} = [\hat{\mathbf{R}}, \hat{\mathbf{T}}]^\top$ is given by solving the normal equation:

$$\hat{\mathbf{x}} = (\mathbf{A}^\top\mathbf{WA} + \lambda\mathbf{I})^{-1}\mathbf{A}^\top\mathbf{Wb}. \tag{6}$$

This operation involves only differentiable matrix multiplication and inversion (2×2), allowing gradients to flow end-to-end. To further suppress the heavy

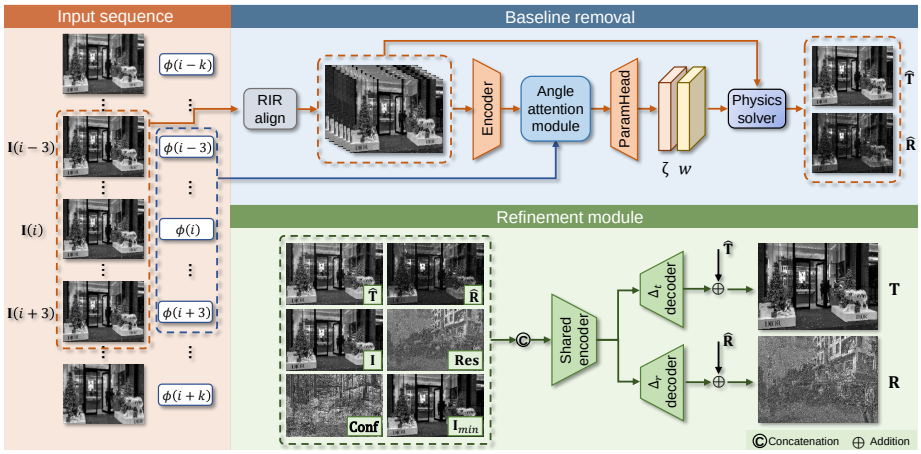

Fig. 4: The two-stage framework inputs a sliding window of mixed spike frames $\mathbf{I}(\cdot)$ and polarization angles $\phi(\cdot)$, and outputs reflection-free videos. In the first stage, baseline removal utilizes an RIR module to align the sequence, followed by an Angle Attention module to enable physically grounded separation of the reflection $\hat{\mathbf{R}}$ and transmission $\hat{\mathbf{T}}$. In the refinement module, the network utilizes physics-discrepancy modeling and a physics-solver confidence map to facilitate the reconstruction of missing high-frequency details. The output reflection and transmission are denoted as $\mathbf{R}$ and $\mathbf{T}$.

noise, we adopt the iterative reweighted least squares (IRLS) strategy. The solver executes over multiple iterations, during which the weight parameters w_k are dynamically updated based on the residual errors from the preceding estimation.

Discussion. RWPS serves as a physically-grounded and computationally efficient solver for initial layer separation. However, while the relative zero-motion observation provides a foundational framework, its idealized nature often leads to the oversmoothing of dynamic scene details, resulting in the loss of critical high-frequency textures. As illustrated in Fig. 3, the reconstructed base transmission layer, consequently lacks fine-grained fidelity. This necessitates a subsequent refinement stage to restore missing structural information and enhance the overall visual quality.

3.3 Refinement module

As discussed in Sec. 3.2, the baseline results $[\hat{\mathbf{R}}, \hat{\mathbf{T}}]$ contain noise and lack fine-grained textural details. To bridge the gap between the analytic physical base and the photorealistic target, we design a refinement module, which leverages physical discrepancy cues and uncertainty-aware spatial modulation. As shown in Fig. 4, the network utilizes a dual-stream architecture conditioned on the physical solver's internal state. It begins by extracting features from the RIR-aligned observation sequence. Since the polarization modulation process is strictly controlled by the polarizer angle, we integrate an angle self-attention module into

the encoder. Specifically, each angle is encoded as $[\cos 2\phi(i), \sin 2\phi(i)]$ to respect the π-periodicity of linear polarization. A two-head attention block projects image features to queries, injects the angle embedding into keys and values, and performs softmax weighting over the $2K+1$ angular observations at each spatial location. The attended feature is added back to the original feature and is used to predict the harmonic mixing parameters and reliability weights. With this design, the encoder can emphasize frames with more discriminative and orthogonal physical states, leading to more accurate results from the physics solver described in Sec. 3.2.

Physics-discrepancy modeling. We map the two layers back into the observation space by applying the estimated mixing coefficient $\hat{\zeta}(i)$:

$$\hat{\mathbf{I}}(i) = \frac{1}{2}\Big(\hat{\zeta}(i)\hat{\mathbf{R}} + \big(1-\hat{\zeta}(i)\big)\hat{\mathbf{T}}\Big). \tag{7}$$

By comparing this physical reconstruction with the actual aggregated spike observation $\mathbf{I}(i)$, we obtain the back-projection residual map: $\mathbf{Res}(i) = \mathbf{I}(i) - \hat{\mathbf{I}}(i)$. This residual map $\mathbf{Res}(i)$ serves as a critical data-fidelity indicator, precisely identifying the high-frequency textures and structural discrepancies introduced by the baseline removal stage.

Physics-solver confidence map. We introduce a confidence map which is an intermediate result of the physics solver. Let $S_{rr} = \sum_k w_k \zeta_k^2$, $S_{tt} = \sum_k w_k (1-\zeta_k)^2$, and $S_{rt} = \sum_k w_k \zeta_k (1-\zeta_k)$. We compute $\mathbf{Conf} \propto S_{rr}S_{tt} - S_{rt}^2$ and normalize it to $[0,1]$. This determinant-like score measures the linear independence of the transmission and reflection components within the observed polarization sequence. In regions with high confidence, the network acts conservatively, preserving the reliable structures from the physics solver; conversely, in noise-corrupted or ill-posed regions, the network dynamically upweights the data-driven pathways to rectify textures.

With this physical guidance, we can construct the base input tensor by concatenating the initial estimates, the raw observation, a multi-angle minimum map $\mathbf{I}_{min}(i)$ (acting as a transmission prior), a solver confidence map and the residual map: $\mathbf{x} = \big[\hat{\mathbf{T}},\ \hat{\mathbf{R}},\ \mathbf{I},\ \mathbf{I}_{min},\ \mathbf{Conf},\ \mathbf{Res}\big]$. Finally, the network predicts the high-frequency residual details (Δ_t, Δ_r) through two coupled decoder branches. The final high-fidelity reflection separation is obtained by adding these details back to the physical base: $\mathbf{T} = \hat{\mathbf{T}} + \Delta_t(\mathbf{x}), \mathbf{R} = \hat{\mathbf{R}} + \Delta_r(\mathbf{x})$. By learning the residuals rather than synthesizing the images from scratch, our network successfully circumvents the semantic ambiguity problem common in pure image-based methods, achieving robust separation even in highly dynamic sequences.

3.4 Loss and training

We train the network end-to-end with a compact objective that enforces (i) center-frame layer fidelity, (ii) structural sharpness, and (iii) physical consistency with the polarization observation model. Let $\mathbf{T}$ and $\mathbf{R}$ denote the predicted transmission and reflection of the center frame in a K-frame window, and $\mathbf{T}^{\mathrm{gt}}, \mathbf{R}^{\mathrm{gt}}$ be their ground truths.

We supervise the center-frame layers using a weighted L_1 loss, with an additional weighted gradient term to preserve sharp edges in the transmission. To suppress reflection residuals, we assign higher weights to regions with strong reflection energy:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{T} - \mathbf{T}^{\text{gt}}\|_{1,\mathbf{M}} + \lambda_r \|\mathbf{R} - \mathbf{R}^{\text{gt}}\|_1, \qquad \mathcal{L}_{\text{grad}} = \|\nabla\mathbf{T} - \nabla\mathbf{T}^{\text{gt}}\|_{1,\mathbf{M}}, \quad (8)$$

where $\|X\|_{1,\mathbf{M}} = \sum_p \mathbf{M}(p)|X(p)|$ and $\mathbf{M} \in [0,1]$ is a soft reflection-dominance mask constructed from the synthetic mixing ratio.

Given the input spike frame stack $\{\mathbf{I}_{\phi_k}\}_{k=1}^K$ and angles $\{\phi_k\}_{k=1}^K$, we enforce that the predicted layers can physically re-render the observations. We synthesize the observations as:

$$\hat{\mathbf{I}}_{\phi_k} = \frac{1}{2}\left(\hat{\zeta}_k \mathbf{R} + (1 - \hat{\zeta}_k)\,\mathbf{T}\right), \quad (9)$$

where $\hat{\zeta}_k$ is the predicted mixing coefficient. We constrain both the physical re-rendering error and the parameter estimation accuracy:

$$\mathcal{L}_{\text{phys}} = \frac{1}{K}\sum_{k=1}^{K}\left\|\hat{\mathbf{I}}_{\phi_k} - \mathbf{I}_{\phi_k}\right\|_1 + \lambda_\zeta \left\|\hat{\zeta} - \zeta^{\text{gt}}\right\|_1. \quad (10)$$

To prevent structural leakage between the decoupled layers, following Zhang *et al*. [37], we introduce the gradient exclusion term. To enforce such spatial decorrelation while being robust to extreme spike noise, we apply a soft-thresholded exclusion loss:

$$\mathcal{L}_{\text{excl}} = \left\| \tanh(\lambda_{ex}|\nabla\mathbf{T}|) \odot \tanh(\lambda_{ex}|\nabla\mathbf{R}|)\right\|_1, \quad (11)$$

where $\odot$ is the element-wise multiplication, and the tanh function, scaled by λ_{ex}, softly saturates excessively large gradients to stabilize training. The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_g \mathcal{L}_{\text{grad}} + \lambda_p \mathcal{L}_{\text{phys}} + \lambda_e \mathcal{L}_{\text{excl}}. \quad (12)$$

We train and evaluate our framework on the proposed synthetic dataset. We implement the model using PyTorch on a single NVIDIA RTX 4090 GPU. Training is conducted with a batch size of 8 and an Adam optimizer with an initial learning rate of 3×10^{-4}. To stabilize the coupling between the RWPS and the neural network, we employ a progressive three-phase curriculum strategy over 60 epochs: (1) a warm-up phase (Epoch 1-5) with strong teacher forcing ($\alpha = 1.0$) for the parameter estimator. Here, α denotes the teacher-forcing ratio that controls the blend between the ground-truth mixing coefficients ζ^{gt} and the predicted ones $\hat{\zeta}$ injected into the physics solver, which effectively prevents gradient explosions caused by early-stage inaccurate predictions; (2) a transition phase (Epoch 6-15) where the reliance on the ground-truth coefficients is reduced ($\alpha = 0.5$), and supervision gradually shifts to the refined output; (3) and a fine-tuning phase (Epoch 16-60) where the network operates entirely autonomously ($\alpha = 0.0$) using only its predicted parameters, enforced by strict physical constraints ($\lambda_{\text{phys}} = 0.1$, $\lambda_{\text{excl}} = 0.01$).

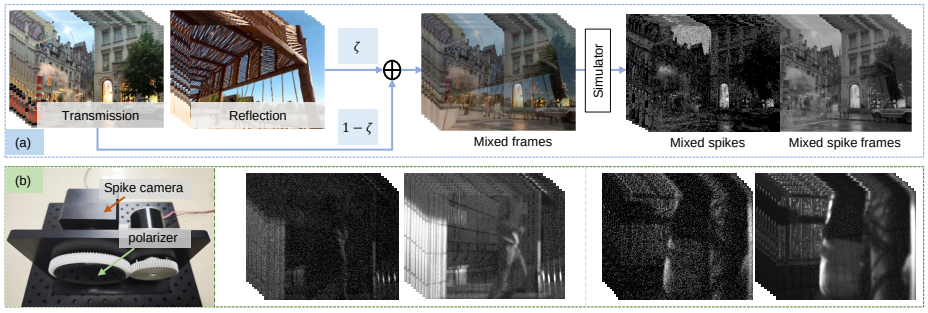

Fig. 5: Overview of the dataset construction pipeline. (a) We mix the video frames from the transmission layer and the reflection layer; the mixed video frames are then simulated to generate mixed spikes, which are then aggregated into spike frames. (b) A spike camera system with a rotating polarizer. Two groups of examples are shown in this figure.

3.5 Dataset

Obtaining pixel-aligned ground truth for high-speed reflection separation is physically intractable in the real world. To train and evaluate our framework, we construct a comprehensive dataset comprising both synthetic data and real-scene data (as shown in Fig. 5).

Synthetic data. To enable supervised learning, we synthesize training samples by implementing the forward observation models described in Sec. 3.1 as our rendering pipeline. We collect video sequences from the Internet to serve as the transmission **T** and reflection **R** video layers. To ensure physical diversity, we continuously sample the virtual polarizer angle $\phi(t)$ covering 180° and simulate spatially varying incidence planes. As shown in Fig. 5 (a), the continuous mixed light is then computationally rendered frame-by-frame according to the Fresnel-based mixing model defined in Eqns. (1) (2). To synthesize spikes, we feed the mixed frames into a spike simulator [8]. In total, we synthesize 50 groups of synthetic data, each comprising one second of continuous temporal data. During training, the 20,000 Hz spike streams are densely sampled with sliding temporal windows, which provides a large set of local training clips from each continuous group. We allocate 40 groups for training and reserve the remaining 10 groups for testing.

Real-scene data. As illustrated in Fig. 5(b), we design the hardware prototype capable of driving a polarizer at high speeds. The polarizing filter is mounted on a gear system and rotates at 30 CPS. This rapid rotation is powered by a high-speed motor coupled with a side-mounted ceramic bearing. For acquisition, the motor first runs for about 3 seconds to reach a stable speed. We calibrate the initial polarizer angle using the darkest response of a polarized white-light smartphone screen, and then infer later angles from the stable time-angle mapping of the continuous spike stream. The spike count within a short gate follows the integrate-and-fire model in Sec. 3.1 and therefore approximates linear inten-

sity, while a very small Δt still reduces SNR. Utilizing this acquisition setup, we captured 15 real dynamic sequences to serve as our testing set.

4 Experiments

We evaluate on both synthetic and real-scene datasets. To validate its effectiveness, we compare our framework against representative state-of-the-art methods: the single-image-based RDNet [38], and two polarization-based approaches, PolarFree [34] and Lyu *et al.* [20]. Since the compared methods are designed for standard frame-based images, the comparison is an adapted reference comparison rather than a same-modality setting. We therefore convert our linear spike frames to their expected input formats while preserving their original assumptions as much as possible. We apply gamma correction ($\gamma = 2.2$) and channel replication to meet the data format requirements of these methods. For polarization methods, we select frames from our aligned-stack with angles closest to their required configurations (*i.e.*, $\{0°, 45°, 90°, 135°\}$ for PolarFree [34]).

4.1 Evaluation on synthetic data

We evaluate the reconstructed reflection-free videos in terms of PSNR, SSIM [31], and LPIPS [36]. The quantitative results listed in Table 1 demonstrate that our method achieves the best performance across all metrics, validating the effectiveness of our specialized physics solver for spike data. Visual comparisons are presented in Fig. 6. As a single-image method, RDNet only leverages limited spatial texture cues and thus fails to handle strong specular highlights and complex reflections, producing obvious residual artifacts. While PolarFree [34] and Lyu *et al.* [20] use polarization to remove reflections, their outputs remain ghosting artifacts and residual noise due to insufficient handling of the low SNR. In contrast, our method recovers accurate transmission structures with rich high-frequency textures, demonstrating superior robustness in the spike domain.

4.2 Evaluation on real-scene data

To validate the effectiveness on real-world scenes, we further evaluate our model on the real-scene dataset. Figure 7 illustrates the qualitative performance. The input frames suffer from low SNR, which is typical for high-speed spike capture through polarizing filters. Under such attenuation, existing methods struggle to effectively remove reflections, as shown in Fig. 7. As shown in the first example of Fig. 7, the swinging sandbag is successfully recovered without being corrupted by reflections. Similarly, in the second row, the fine-grained details of the human subject are faithfully restored. In the third and fourth groups, the reflections are nearly global, casting a severe haze over the entire transmission layer. Compared to the baselines, our approach effectively removes these reflections and recovers clear and high-fidelity transmissions.

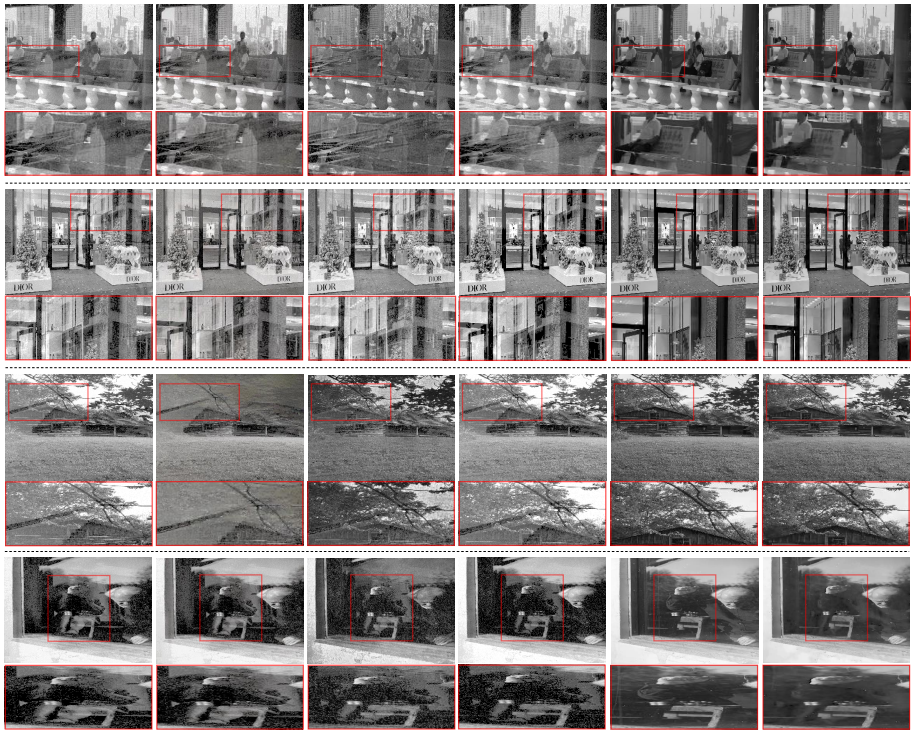

Fig. 6: Qualitative comparisons on the synthetic dataset. Please zoom in for more details. The corresponding videos are available in the supplementary materials.

It is worth noting that while the baseline results in Fig. 7 successfully separate a portion of the reflections in terms of physical structure, the recovered transmission layer loses some high-frequency details. This is caused by the inherent ill-conditioning of polarization modulation regions and the ridge regularization adopted to suppress extreme noise, leading to an over-smoothed visual appearance. Our results not only eliminate most interfering reflections but also reconstruct sharp edges and faithful high-frequency textures of the transmission layer, without introducing extra noise amplification or artifacts. This demonstrates that our method is robust to real-world sensor imperfections and can generalize well beyond synthetic data.

4.3 Ablation study

We validate the contribution of our proposed modules through a series of ablation studies. As shown in Table 1, removing the RIR module leads to a notable performance degradation. This is mainly because the inherent flicker and slight-motion introduced by the rotating polarizer compromise the pixel-level align-

Fig. 7: Qualitative comparison results on real captured scene data. Our proposed method successfully removes the reflection while preserving transmission details even under complicated real-world noise conditions. The full comparative video sequences for visual comparison are available in the supplementary materials.

ment requirements of the physical observation model, resulting in artifacts and background blurring in Fig. 8. Second, the angle attention module, by embedding polarization angle information, enables the network to effectively correlate light intensity fluctuations with polarization states. Disabling this module leads to an observable drop in reconstruction performance and makes it difficult to decouple mixed signals in the weak polarization modulation region. Finally, the solver confidence map provides guidance on spatial uncertainty in optimization. As shown in Table 1, the lack of confidence guidance also leads to a noticeable performance degradation. Beyond module ablation, the supplementary material reports reduced-angle/window, motion-speed, and solver analyses.

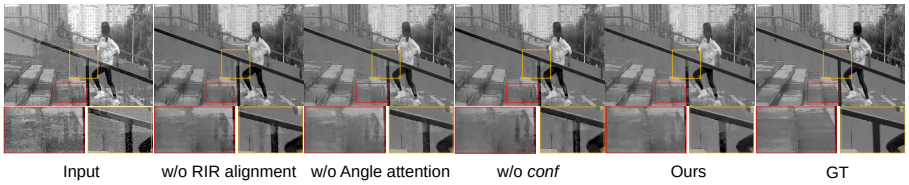

Fig. 8: Qualitative ablation results on synthetic datasets. As observed, removing the key modules leads to a noticeable degradation in reconstruction quality.

Table 1: Quantitative measurements of comparison with state-of-the-art methods and the ablation study. ↑ (↓) indicates larger (smaller) values are better. Bold numbers indicate the best results.

Metric	Comparison with state-of-the-art methods				Ablation studies		
	RDNet [38]	PolarFree [34]	Lyu *et al.* [20]	Ours	w/o RIR alignment	w/o Angle attention	w/o *conf*
PSNR↑	17.4130	16.3870	16.5347	**27.0598**	26.7932	26.9651	26.8910
SSIM↑	0.6790	0.5216	0.7010	**0.8638**	0.8436	0.8553	0.8470
LPIPS↓	0.3072	0.4692	0.2658	**0.0894**	0.1111	0.0938	0.1012

5 Conclusion

In this paper, through a comprehensive analysis of polarization-based reflection removal, we identified that continuously tracking dense polarization states using a high-speed spike camera can resolve the temporal limitations of conventional cameras. In dynamic scenes, compared to standard frame-based acquisition that suffers from discrete temporal sampling, our rapidly rotating polarizer setup successfully captures continuous polarization states. Accordingly, we develop an effective physically-driven framework to achieve 340 FPS reflection-free video reconstruction. Under the joint action of the proposed RWPS and the physics-discrepancy guided refinement module, our framework demonstrates the effectiveness on both synthetic and real-scene data.

Limitations. While our framework achieves robust reflection-free video reconstruction in dynamic scenes, the light attenuation limits its deployment in low-light environments. The relative zero-motion observation can also break down when the accumulated displacement over the observation window becomes too large. Future work will explore computational exposure strategies, adaptive window selection, and non-linear mixing models to overcome these physical limitations.

Acknowledgements

This work was supported by National Key R&D Program of China (2024YFB3909900), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), Beijing Natural Science Foundation (Grant No. L233024), and Beijing High Innovation Plan (Contract No. 202604841443).

References

1. PolarCam data sheet. https://www.4dtechnology.com/wp-content/uploads/Data_Sheets/PolarCam_Data_Sheet.pdf (2018), accessed: 2026-06-25
2. Chang, Y., Jung, C.: Single image reflection removal using convolutional neural networks. IEEE Transactions on Image Processing **28**(4), 1954–1966 (2019)
3. Chang, Y., Jung, C., Sun, J., Wang, F.: Siamese dense network for reflection removal with flash and no-flash image pairs. International Journal of Computer Vision **128**(6), 1673–1698 (2020)
4. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proc. of International Conference on Computer Vision. pp. 5017–5026 (2021)
5. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: Proc. of International Conference on Computer Vision. pp. 3258–3267 (2017)
6. Fang, S., Wang, Y., Zhang, J., Li, Z., Wang, Y.: Adaptive language-aware image reflection removal network. In: Proc. of International Joint Conference on Artificial Intelligence. pp. 973–981 (2025)
7. 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)
8. Hu, L., Ma, L., Guo, Y., Huang, T.: Scsim: A realistic spike cameras simulator. In: IEEE International Conference on Multimedia and Expo. pp. 1–6. IEEE (2024)
9. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: Proc. of International Conference on Computer Vision. pp. 13138–13147 (2023)
10. Huang, T., Zheng, Y., Yu, Z., Chen, R., Li, Y., Xiong, R., Ma, L., Zhao, J., Dong, S., Zhu, L., et al.: 1000× faster camera and machine vision with ordinary devices. Engineering **25**, 110–119 (2023)
11. Kong, N., Tai, Y.W., Shin, J.S.: A physically-based approach to reflection separation: from physical modeling to constrained optimization. IEEE transactions on pattern analysis and machine intelligence **36**(2), 209–221 (2014)
12. Kong, N., Tai, Y.W., Shin, S.Y.: A physically-based approach to reflection separation. In: Proc. of Computer Vision and Pattern Recognition. pp. 9–16 (2012)
13. Lei, C., Chen, Q.: Robust reflection removal with reflection-free flash-only cues. In: Proc. of Computer Vision and Pattern Recognition. pp. 14811–14820 (2021)
14. Lei, C., Huang, X., Zhang, M., Yan, Q., Sun, W., Chen, Q.: Polarized reflection removal with perfect alignment in the wild. In: Proc. of Computer Vision and Pattern Recognition (2020)
15. Levin, A., Weiss, Y.: User assisted separation of reflections from a single image using a sparsity prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(9), 1647–1654 (2007)
16. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proc. of Computer Vision and Pattern Recognition. pp. 3565–3574 (2020)
17. Li, Y., Brown, M.S.: Single image layer separation using relative smoothness. In: Proc. of Computer Vision and Pattern Recognition. pp. 2752–2759 (2014)
18. Li, Y., Liu, M., Yi, Y., Chen, Q., Zuo, W.: Robustsirr: Robust single image reflection removal with transformer. In: Proc. of Computer Vision and Pattern Recognition. pp. 5696–5705 (2022)

19. Li, Y., Liu, M., Yi, Y., Li, Q., Ren, D., Zuo, W.: Two-stage single image reflection removal with reflection-aware guidance. Applied Intelligence **53**, 19433–19448 (2023)
20. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Reflection separation using a pair of unpolarized and polarized images. In: Proc. of Advances in Neural Information Processing Systems (2019)
21. 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 (2022)
22. Nayar, S.K., Fang, X.S., Boult, T.: Separation of reflection components using color and polarization. International Journal of Computer Vision **21**(3), 163–186 (1997)
23. Prasad, B.H.P., S, G.R.K., Boregowda, L.R., Mitra, K.: Burst reflection removal using reflection motion aggregation cues. In: Proc. of Winter Conference on Applications of Computer Vision. pp. 239–248 (2023)
24. Punnappurath, A., Brown, M.S.: Reflection removal using a dual-pixel sensor. In: Proc. of Computer Vision and Pattern Recognition. pp. 1556–1565 (2019)
25. Schechner, Y.Y., Kiryati, N., Basri, R.: Separation of transparent layers using focus. International Journal of Computer Vision **39**(1), 25–39 (2000)
26. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: Proc. of Computer Vision and Pattern Recognition. pp. 3193–3201 (2015)
27. Song, Z., Zhang, Z., Zhang, K., Luo, W., Fan, Z., Ren, W., Lu, J.: Robust single image reflection removal against adversarial attacks. In: Proc. of Computer Vision and Pattern Recognition. pp. 24688–24698 (2023)
28. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: CRRN: Multi-scale guided concurrent reflection removal network. In: Proc. of Computer Vision and Pattern Recognition. pp. 4777–4785 (2018)
29. Wan, R., Shi, B., Hwee, T.A., Kot, A.C.: Depth of field guided reflection removal. In: Proc. of International Conference on Image Processing. pp. 21–25 (2016)
30. Wan, R., Shi, B., Li, H., Duan, L.Y., Tan, A.H., Chichung, A.K.: CoRRN: Cooperative reflection removal network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
31. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers. vol. 2, pp. 1398–1402 (2003)
32. Wieschollek, P., Gallo, O., Gu, J., Kautz, J.: Separating reflection and transmission images in the wild. In: Proc. of European Conference on Computer Vision. pp. 89–104 (2018)
33. Yang, J., Li, H., Dai, Y., Tan, R.T.: Robust optical flow estimation of double-layer images under transparency or reflection. In: Proc. of Computer Vision and Pattern Recognition. pp. 1410–1419 (2016)
34. Yao, M., Wang, M., Tam, K.M., Li, L., Xue, T., Gu, J.: Polarfree: Polarization-based reflection-free imaging. In: Proc. of Computer Vision and Pattern Recognition. pp. 10890–10899 (2025)
35. Zakarin, D., Wandel, T., Obukhov, A., Dai, D.: Reflection removal through efficient adaptation of diffusion transformers. In: Proc. of Computer Vision and Pattern Recognition. pp. 2776–2785 (2026)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of Computer Vision and Pattern Recognition. pp. 586–595 (2018)

37. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proc. of Computer Vision and Pattern Recognition. pp. 4786–4794 (2018)
38. Zhao, H., Li, M., Hu, Q., Guo, X.: Reversible decoupling network for single image reflection removal. In: Proc. of Computer Vision and Pattern Recognition. pp. 26430–26439 (2025)
39. Zheng, Q., Shi, B., Chen, J., Jiang, X., Duan, L.Y., Kot, A.C.: Single image reflection removal with absorption effect. In: Proc. of Computer Vision and Pattern Recognition (2021)
40. Zhong, H., Hong, Y., Weng, S., Liang, J., Shi, B.: Language-guided image reflection separation. In: Proc. of Computer Vision and Pattern Recognition. pp. 24913–24922 (2024)