Flash-Split: 2D Reflection Removal with Flash Cues and Latent Diffusion Separation

Tianfu Wang* Mingyang Xie*† Haoming Cai Sachin Shah Christopher A. Metzler

University of Maryland

mingyang@umd.edu

Abstract

Transparent surfaces, such as glass, create complex reflections that obscure images and challenge downstream computer vision applications. We introduce Flash-Split, a robust framework for separating transmitted and reflected light using a single (potentially misaligned) pair of flash/no-flash images. Our core idea is to perform flash-informed reflection separation iteratively in a low-dimensional latent space. Specifically, Flash-Split consists of two stages. Stage 1 separates the reflection latent and transmission latent via a dual-branch diffusion model that is conditioned on an encoded flash/no-flash latent pair. This stage effectively mitigates the flash/no-flash misalignment issue. Stage 2 restores high-resolution, faithful details to the separated latents via a cross-latent decoding process that is conditioned on the original images before separation. We validate Flash-Split on challenging real-world scenes and demonstrate it significantly outperforms existing methods. This paper appears at CVPR 2025.

1. Introduction

Scenes with transparent surfaces, especially glass, frequently surround us and create specular reflections. In such scenes, what we perceive is a combination of transmitted and reflected light. This study aims to separate the transmitted and reflected 2D scenes.

Reflection removal and separation have garnered significant interest in the low-level vision community [29, 30, 33, 61, 69]. By removing and separating the reflections in the scene, we can not only enhance visual quality but also boost the performance of downstream vision tasks such as depth estimation, robot navigation, object classification, and scene understanding [22, 53, 56].

Separating the reflection from the transmission is a challenging task due to its highly under-determined nature. With both the transmission and reflection being unknown, it is challenging to solve for each of them just based on their summed intensities. To overcome this challenge, existing approaches have leveraged various prior assumptions. Some approaches, for example, assume the reflection is out of focus [3, 64] or that the front and back sides of the glass cause significant double reflection [45]. However, these assumptions might not always hold in real-world scenarios.

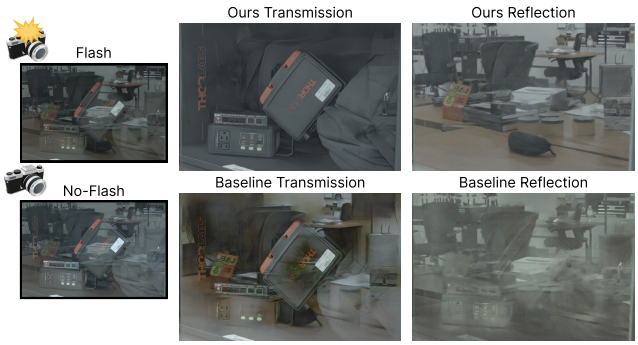

Figure 1. **Left**: We separated the transmitted and reflected scenes by capturing one image with camera flash and another with no flash, despite them being potentially misaligned due to hand shake. **Right**: Our proposed Flash-Split method archives a precise separation of the transmission and the reflection, performing much better than the baseline [31].

On the other hand, adding illumination control, e.g., using the built-in flash of a camera, is both accessible to everyday users and demonstrates significant potential in reducing the under-determined nature of reflection separation. Specifically, the flash/no-flash technique [1] performs reflection separation by capturing two images from the same viewpoint: one image with the camera flash on and another image with the camera flash off. While the camera flash boosts the intensity of the transmitted scene, it mostly leaves the reflected scene unchanged (more details in Section 3.2). Consequently, by subtracting the image captured without flash from the image captured with flash, we can retrieve a transmission scene free of reflections [29].

*Equal contribution.

†Corresponding author: mingyang@umd.edu

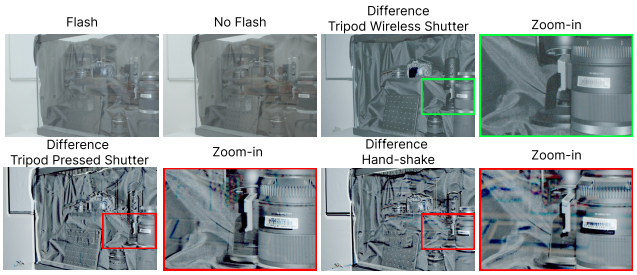

Figure 2. **Conventional Flash/No-Flash Methods Need Perfectly Paired Captures**. The camera flash increases the brightness of the transmitted scene without affecting that of the reflected scene. Therefore, the difference between this pair will be the transmitted scene free of reflection. **Top Right**: If we capture a perfectly aligned pair of flash/no-flash images using a tripod plus wireless shutter control, the difference is a perfect transmission image. **Bottom Left**: if we use a tripod but use a finger to press the shutter button, this slight motion will cause the two shots to be misaligned from each other, leading to noticeable artifacts in the difference image. **Bottom Right**: if we just do handheld photography, the difference image exhibits even stronger artifacts. **Takeaway**: this misalignment issue has been the key barrier to applying flash/no-flash photography, an accessible method with great potential, to the task of reflection removal. In our work, we propose a robust approach to circumvent this key barrier.

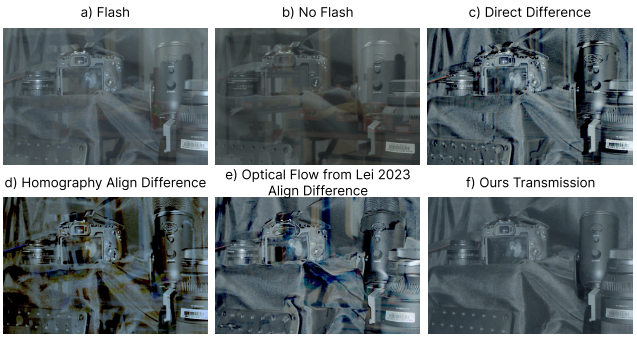

Figure 3. **Aligning Flash/No-Flash Images Is A Difficult Task for Image Registration Methods**. While the difference between a misaligned flash/no-flash image pair (a,b) exhibits severe artifacts (c), aligning them is a non-trivial problem, since camera flash modifies the appearance of the transmitted component of one of the two images. Existing registration methods, like homography (d [12]) or optical flow prediction (e [47]) used in Lei et al. [31], fail to align this pair of images well – their aligned flash/no-flash pair still suffer from severe artifacts. In contrast, our method (f) circumvents the misalignment issue by directly encoding the flash/no-flash pair into the latent space to perform iterative latent diffusion separation, eventually yielding a clean transmission scene.

The primary limitation of this approach is its reliance on precisely aligned flash/no-flash image captures, meaning the camera must remain stationary between shots. Even minimal movements, like pressing the shutter, can misalign the image pair, rendering this approach ineffective (Fig. 2). To use this approach, users either need to hold their hands perfectly still during the two-shot capture, which is realistically infeasible for humans, or use a tripod plus remote shutter control. This requirement for paired captures poses a significant challenge for effective reflection separation in uncontrolled, real-world conditions.

To overcome this limitation, Lei et al. [31] explored pre-align the flash image and the no-flash image via an optical flow module [47]. However, aligning the flash and no-flash images is much more difficult than the usual optical flow/homography task since the flash modifies the appearance of the objects in the transmitted scene. From our empirical experiments, we found that such pre-alignment is not robust when evaluating on diverse real-world flash/no-flash images (Fig. 3).

In our paper, we develop a novel approach for robust reflection separation, using a pair of misaligned flash/no-flash images. The key idea in this paper is to leverage flash cues to perform *latent-space* reflection separation. Our intuition behind it is that the condensed latent space makes it easier for our model to perform reflection separation under the flash guidance, while being more robust to the flash/no-flash misalignment issue. Guided by this idea, we further decouple the reflection separation problem into two consecutive stages: (1) iterative latent diffusion separation and (2) cross-latent decoding.

In Stage 1 of our method, given a potentially misaligned flash/no-flash image pair as input, we first encode it into a flash/no-flash latent pair using a VAE encoder [26]; afterward, we iteratively separate apart the latent representations for the reflected scene and the transmitted scene, using a dual-branch diffusion model conditioned on the flash/no-flash latent pair. By implicitly leveraging the flash cues at latent space, our dual-branch diffusion model learns to effectively distinguish between the features from the transmitted scene and those from the reflected scene.

However, while we can effectively separate apart the transition and reflection in the latent space, naively decoding them to RGB space will tend to hallucinate content details (especially the high-frequency details) due to the inevitable under-determinedness of the decoding process, therefore reducing the faithfulness of the final separated images. To solve this, we introduce our Stage 2 cross-latent decoding, where we use the separated latent as guidance to extract the sharp image features from the unseparated input image. On a higher level, we are fusing the well-separated yet highly condensed information in our predicted latents with the unseparated yet highly detailed input image, to reconstruct a well-separated image that preserves fine and faithful details. By combining Stage 1 and Stage 2, our proposed method significantly outperforms baselines and has been validated on challenging real-

world scenes, even when being trained from scratch. Additionally, we show that our performance can be further improved by leveraging weights of pretrained latent diffusion models [43].

Our contributions are:

- We propose Flash-Split, a robust 2D reflection removal framework that combines flash/no-flash physical cues and latent space transmission/reflection separation.
- We develop a dual-branch diffusion framework for iterative latent diffusion separation, which can effectively handle misaligned flash/no-flash input images.
- We use a cross-latent decoding module to restore faithful and high-frequency details from our separated latents.
- We demonstrate that our approach effectively separates reflection and transmission in real scenes, outperforming all baselines, including other flash/no-flash-based methods, in challenging cases with strong reflections.

2. Related Works

Reflection removal has been a long-standing task in computational photography, with existing methods roughly categorized into three general categories: software-only, multi-view, and hardware-based.

Software-only Reflection removal. The majority of software-only reflection removal works take only one single image with mixed transmission and reflection components and attempt to separate the two. Traditionally, this involves using prior statistical properties of the reflection [32–34, 37]. Recently, deep learning methods [3, 9, 10, 16, 20, 21, 24, 25, 35, 38, 45, 48–51, 55, 57, 58, 63, 64, 67–70] have emerged where the reflection can be learned in a data-driven manner with promising results. Nevertheless, given the inherent ill-posed nature of reflection removal, software-only approaches are not robust to complex real-world scenarios, especially in scenes with strong reflections.

Multi-View Reflection Removal. Multi-frame approaches [2, 8, 11, 13–15, 18, 19, 37, 39, 61, 62] aims to combine temporal and spatial cues for consistent reconstruction and separation. Among them, unsupervised works such as NeRFRen [15] use neural fields to model both the transmitted and reflected 3D scenes, leveraging cross-view consistency as cues for 3D reflection separation.

Hardware-Related Reflection Removal. These methods introduce hardware elements to exploit optical cues of the transmission and reflection light transport. Some studies employ polarization cues [27, 28, 30, 36, 40, 41]. They leverage the fact that the transmission is unpolarized while the reflection component varies when rotating the polarization filter. However, access to polarization cameras is limited to general camera/smartphone users. Other works, including our paper, involve taking a pair of flash/no-flash images from the same view point [29, 59, 60], the mechanism of which will be introduced in detail in the next section.

3. Proposed Method

3.1. Flash/No-Flash Preliminaries

An established technique [29] among photographers to obtain a reflection-free image is to compare images taken with and without flash from the same viewpoint. Assume that we have a composite scene consisting of a transmission-only scene $\mathbf{T}$, a reflection-only scene $\mathbf{R}$, and a transparent reflective surface, e.g., glass. The image of this composite scene $\mathbf{I}$ can be formulated as

$$\mathbf{I} = \mathbf{T} + \gamma \circ \mathbf{R} \tag{1}$$

Now, assume we take a second image from the exact same viewpoint as the first image, only now turning on an additional illumination source co-located on the viewpoint, e.g., a camera flash. Assuming the illumination (camera flash) strength is uniformly distributed, it will increase the intensity across all pixels in the transmission scene in proportion to each pixel's reflectivity. We further discuss the applicability of flash/no-flash photography in Sec. 15 of the Suppl.

Under these conditions, we can now approximate the image taken with flash $\mathbf{I}_{Flash}$ as

$$\mathbf{I}_{Flash} \approx (1 + \theta)\mathbf{T} + \gamma \circ \mathbf{R}, \tag{2}$$

Taking the difference between the flash image $\mathbf{I}_{Flash}$ and no-flash image $\mathbf{I}$, we shall obtain an image of the transmitted scene:

$$\mathbf{I}_{Flash} - \mathbf{I} \approx \theta \circ \mathbf{T}. \tag{3}$$

A visual example of how this approach works is shown in Fig. 2 top row. Note that this difference image will slightly differ from the transmitted scene when there is no glass, because the intensity of the difference image is dependent on the flash illumination strength.

3.2. Our Core Idea

While the flash/no-flash approach has great potential for reflection removal, it requires a *perfectly aligned* pair of images, e.g., from the same camera viewpoint. Otherwise, the flash/no-flash difference will contain heavy artifacts. As shown in the bottom two rows of Fig. 2, any motion during capture, including user hand shake or even pressing the shutter button, will cause this method to break.

To overcome the misalignment issue, the most straightforward way is to align the image pair first before taking the difference, which has been explored by Lei et al. [31]. As shown in Fig. 3, the difference image after alignment by homography [12] or optical flow [47] still suffers from noticeable artifacts, not only because the flash/no-flash method is very sensitive to alignment error, but also because the flash changes the appearance of the transmitted scene, making it harder for registration methods to work.

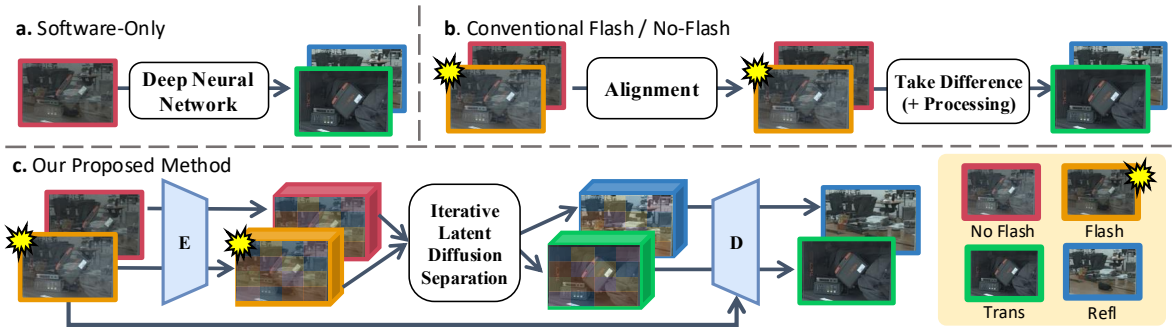

Figure 4. **Comparing Different 2D Reflection Removal Paradigms.** (a): **Software-only methods** pass a single composite image (with both transmission and reflection) to a deep neural net for reflection separation. (b): **Conventional flash/no-flash methods** take the difference of a flash/no-flash image pair to get the transmission image [1]; optionally, one can also use a neural net [7] to predict the reflection image and further refine the transmission image quality (omitted in the figure for simplicity). In cases of misalignment (when not using a tripod), Lei et al. [31] uses an optical flow module to pre-align the image pair. (c): **Our proposed method** encodes the flash/no-flash method down to the latent space: we first encode the flash/no-flash image pair into a flash/no-flash latent pair, then use its physical cue to separate the composite scene's latent into a transmission latent and a reflection latent, and finally decode them back to RGB image space to obtain the clean transmission image and reflection image.

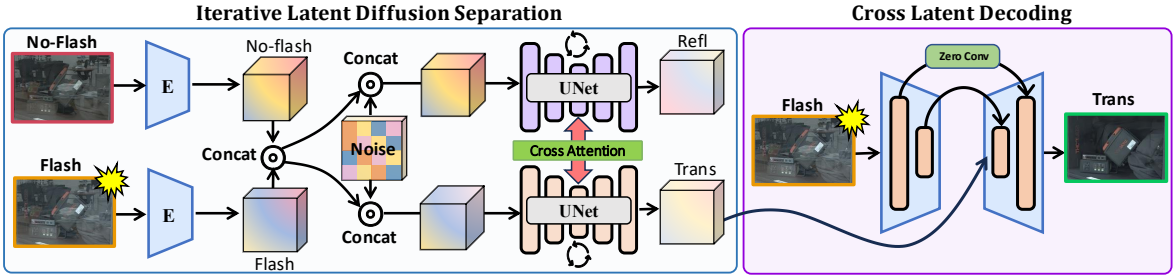

Figure 5. **Our Proposed Pipeline** consists of a latent separation stage and a decoding stage. **Left:** We first encode the misaligned flash/no-flash image pair into a flash/no-flash latent pair. We then use a dual-branch attention UNet with cross-attention in-between to perform latent separation — the goal is to predict a latent for the transmission scene and another latent for the reflection scene. Following recent development of latent diffusion models [23, 43], at each inference step, we concatenate both the flash/no-flash latents with random Gaussian noise and let the dual-branch UNet denoise them. Eventually, the top and bottom branches predict a transmission and reflection latent, respectively. **Right:** We observe that the vanilla decoding process may lead to hallucination and blurriness (Figure 15). To fix this issue, we apply a cross-latent decoding process with a UNet [44] architecture. But unlike a normal UNet, we do not feed the encoder's output into the decoder. Instead, we (1) feed the original unseparated image into the encoder and (2) feed our separated latent (from the first stage) directly into the decoder. The encoder passes information to the decoder only through the skip connection layers. This decoding process combines two complementary sources of information: the predicted latent from Stage 1, separated but missing high-frequency information, and the captured image, unseparated but contains high-frequency details, leading to a faithful reconstruction of the original transmission/reflection scenes.

In our work, we take an alternative path to deal with the misalignment issue. Inspired by recent works on image latent features [26, 42, 43], we take the flash/no-flash method down to the latent space. Similar to how conventional flash/no-flash methods [1] take advantage of a flash/no-flash image pair, we create a flash/no-flash *latent pair* via a variational autoencoder (VAE) [26]. Then, we perform latent-space reflection separation using the flash/no-flash pair to obtain latents for the transmission and reflection images. In the end, we decode the separated latents back to RGB space to obtain the transmission/reflection images. Below we explain why we choose to do separation in latent space.

3.2.1. Latent Separation Mitigates Misalignment

When a vision encoder [26] encodes an image, it expands the feature dimension while reducing the spatial dimension. By focusing on the overall high-level features rather than precise pixel locations, the encoder effectively reduces the impact of spatial misalignment between the flash/no-flash image pair. Specifically, when performing feature extraction, the encoder also aggregates local pixel information, averaging out the difference in misaligned pixel location. We show in Sec. 10 in the Suppl that our method performs robustly against small to moderate misalignment.

3.2.2. Reflection Separation Is Easier in Latent Space

Our key intuition here is that training a model to separate the composite scene's latent into the transmitted scene's latent and the reflected scene's latent will be much easier than training a model to separate a composite image into a transmission image and a reflection image.

More specifically, the high-level representation of image latents allows a model to better focus on separating the main features in the reflected and transmitted scenes(such as the primary object structures). Training with reduced dimensionality also lets a model converge to a better local optimum and have better generalization ability. Once the separation is done in latent space, we then use our customized decoder to restore the fine details, reconstructing a sharp and well-separated image (Further discussed in Sec. 3.3.2).

3.2.3. Leveraging Flash Cues in The Latent Space

Similar to how the flash/no-flash technique works in the RGB image space (Sec. 3.1), after we encode the flash/no-flash image pair into a latent pair, it can still provide important cues for reflection separation, despite being at a condensed latent space. This is because the physical cues from the flash/no-flash technique lie in a relatively low-frequency domain. Intuitively, suppose we have a flash/no-flash image pair with the difference being the transmission; if we downsample them to a smaller dimension, their difference image, despite being low-resolution, would still resemble the transmission scene, serving as a powerful cue.

An ablation is shown in Fig. 16. While keeping the reflection separation happening in latent space, we compare two model variants: one takes flash/no-flash as input, and the other takes in a single image as input. It turns out the former significantly outperforms the latter, implying that the flash/no-flash cues can still be leveraged in latent space.

3.3. Our Pipeline

Following our core idea described in the previous section, we decouple the reflection separation problem into two subproblems: (1) after we first encode the composite (flash/no-flash) images into latents, how to separate them into a latent representation for the transmission image and the reflection image, respectively; and (2): how to restore high-frequency details to the separated latents while keeping the details faithful to the original scene. To address the two issues, we propose a 2 stage framework consisting of iterative latent diffusion separation and cross-latent decoding, described in Sec. 3.3.1 and 3.3.2, respectively. To better illustrate our proposed method, Fig. 4 highlights a high-level comparison between ours and previous works, while Fig. 5 shows our detailed pipeline.

3.3.1. Stage 1: Iterative Latent Diffusion Separation

For latent separation, our method utilizes the iterative latent diffusion process, inspired by recent works [22, 23, 43, 54] For the diffusion denoiser, we developed a dual-branch UNet to jointly predict the transmission latent and the reflection latent, respectively. As shown in Fig. 5, both branches of the UNet are conditioned on the *flash/no-flash latent pair*, which are fixed during the entire diffusion process. Notably, latent diffusion separation also allows us to leverage pretrained large-scale diffusion models [43] as priors for natural images, which can serve as additional guidance for our model's separation.

Furthermore, we place cross-attention [5, 43, 52] between the two branches to iteratively exchange information, Our motivation here is that, while the flash/no-flash pair serve as a strong cue for transmission, it does not provide cue for reflection, which makes the reflection branch more difficult to train. The cross-attention injects the information from the transmission branch to the reflection branch, so that the reflection branch can indirectly leverage the flash cues to improve its prediction, according to the complementary relationship between reflection and transmission. In return, better-predicted reflection can also help with transmission prediction.

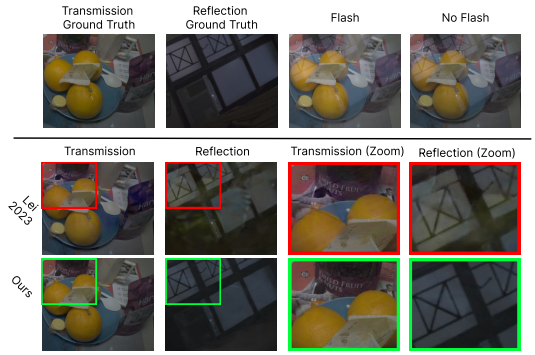

Figure 6. **Qualitative Comparison With Ground Truth Transmission/Reflection.** We qualitatively compare our method with Lei et al. [31] on the dataset introduced by themselves, which contains the ground truth transmission captured by removing the glass that causes the reflection. While our model is trained on exactly the same datasets as theirs, our model performs better reflection separation, due to our latent separation strategy. In addition, note that the reflection strength of scenes in this dataset is very weak, making it easy to separate.

3.3.2. Stage 2: Cross-Latent Decoding

For restoring high-frequency details, the most naive approach is to use a pre-trained VAE decoder [26]; however, we found that the images decoded by it suffer from blurriness, and more importantly, hallucinations (Fig. 15). Given that decoding latent to RGB space is a very underdetermined problem, hallucinations are inevitable unless we supervise the decoding process with other conditions. We notice that the original captured image forms a complementary pair with the separated latent from Stage 1: one is unseparated but contains high-frequency features, and one is well-separated but missing high-frequency. As such, we perform *cross-latent decoding*, as illustrated in Fig. 5.

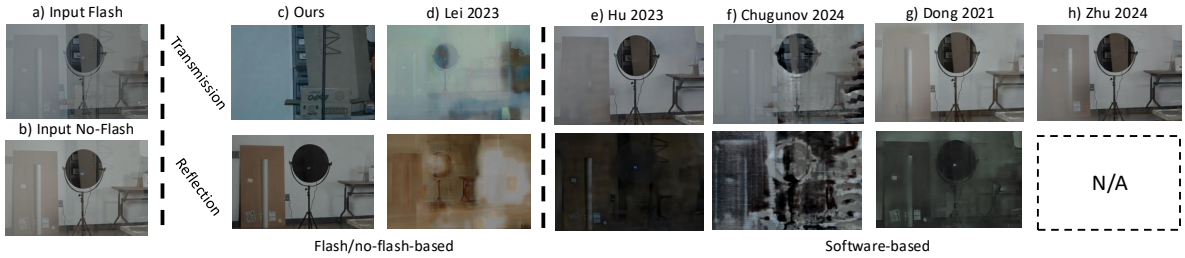

Figure 7. **Real Experiment: The Lab Scene.** The transmission is some paper boxes; the reflection is a door and lamp. Using the top-row flash/no-flash image pair (misaligned due to motion between shots), our method overcomes the misalignment and achieves reflection separation not only better than software-only approaches (e,f,g,h) [8, 9, 21, 69], but also better than another flash/no-flash based method (d) [31], which is trained on exactly the same dataset as ours. Note that Zhu et al. [69] can only predict the transmission, thus the "N/A".

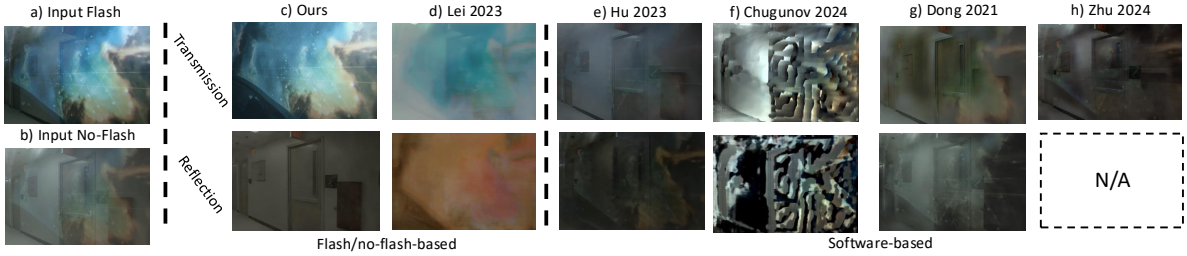

Figure 8. **Real Experiment: The Poster Scene.** The transmission is a poster; the reflection is a hallway.

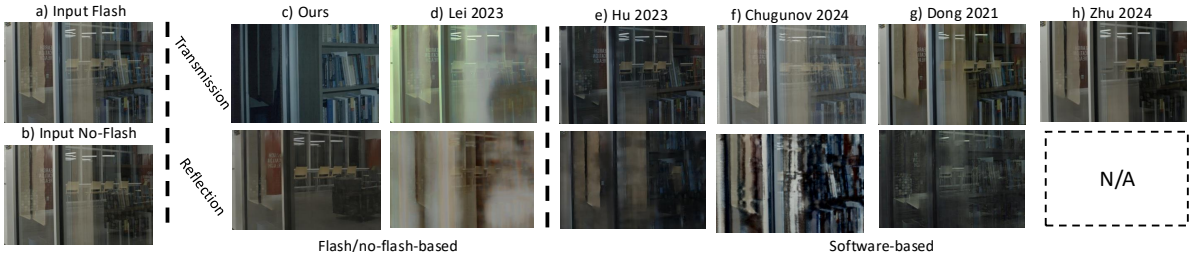

Figure 9. **Real experiment: The Office Scene.** The transmission is a bookshelf behind an office window; the reflection is a study area with chairs and sofas.

Figure 10. **Real experiment: the Outdoor Scene.** The transmission is a toy inside a window; the reflection is an outdoor bench.

.

More specifically, we modify the pre-trained VAE structure to resemble a UNet with zero convolution skip connection but no mid-blocks. We feed the captured image to the encoder and the separated latent into the decoder. The zero convolution facilitates stable training and ensures that the trained decoder does not deviate from the original decoder, which contains rich prior information [65].

With the latent separation and cross-latent decoding stages in place, we now describe the inference procedure for our complete pipeline. We first encode input flash/no-flash images to obtain input latents, and then concatenate them with a noise latent, before passing through the dual-branch diffusion process to iteratively separate apart the latent representations for the transmitted and reflected scenes. Once Stage 1 is finished, the separated latents are then passed through our cross-latent decoders with skip connection guidance from unseparated input images, resulting in clear, separated transmission/reflection images.

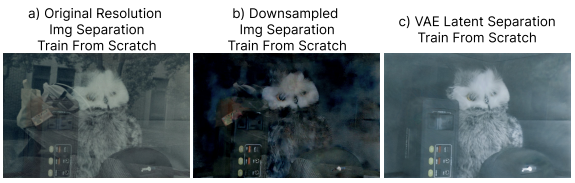

Figure 11. **Latent Space Separation Analysis**. To evaluate the effectiveness of latent diffusion for reflection separation, we train three model variants from scratch: (a) a model without the VAE, directly predicting transmission/reflection at full resolution; (b) a model replacing the VAE with a simple $8\times$ downsampling to match the VAE spatial resolution; and (c) a model incorporating the VAE. Experiments show that the model using VAE achieves the best performance on real-world flash/no-flash images.

Figure 12. **Pretraining Analysis**. We study the effect of pretraining by training our diffusion pipeline from scratch (b) vs. from Stable Diffusion (SD) [43] (c). In real-world testing, although the model from scratch (b) exhibits some minor artifacts compared to the model training pretrained from SD (c), it still performs better than the baseline method [31] in (a).

4. Analysis of Latent Diffusion Separation

Our latent space diffusion separation method improves reflection removal by combining iterative diffusion with compact, high-level representations and leveraging pretrained image priors [43]. In this section we analyze the importance of latent space separation and pretraining.

Importance of Latent Space Separation We evaluate three model variants for reflection removal, including two without a VAE where the diffusion UNet operates in RGB space. The first processes full-resolution images, while the second downsamples inputs by a factor of 8, matching the VAE latent space dimensions. We also train a model using the VAE from scratch. For a fair comparison, all models were trained on $\sim$8K synthetic flash/no-flash images from [31]. Fig. 11 illustrates predicted transmissions on real images from [61] (see Fig. 10 for reference). The full resolution model (Fig. 11a) struggles with reflection separation, indicating that downsampling helps mitigate flash/no-flash misalignment. However, the model using VAE latents (Fig. 11c) outperforms the downsampled variant (Fig. 11b), highlighting the importance of VAE features for effective reflection removal in misaligned images.

Importance of Pretraining While our default model was fine-tuned from pretrained Stable Diffusion (SD) [43], we experimented with retraining it from scratch using only the synthetic flash/no-flash dataset proposed by [31]. Surprisingly, as shown in Fig. 12, the model trained from scratch (Fig. 12b) achieves performance only marginally worse than the model using pretrained SD (Fig. 12c). Notably, the model trained from scratch (Fig. 12b) still performs better than all baseline methods (see Fig. 12a & 10). We present more ablation studies on pretraining in Sec. 14 of the Suppl.

Conclusion Based on our analysis, we believe that pretraining can boost our performance, but is not the dominant reason for our huge advantage over baselines: The iterative diffusion process in the latent space is a much more crucial factor in the effectiveness of our method. Additional experimental results are shown in Sec. 13 of the Suppl.

	Transmission			Reflection		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Dong et al. [9]	25.44	0.907	0.122	23.63	0.803	0.513
Hu et al. [21]	25.61	0.915	0.099	22.44	0.818	0.396
Zhu et al. [69]	25.70	0.919	0.100	–	–	–
SDN [6]	23.44	0.873	0.159	–	–	–
Lei et al. no align [31]	25.41	0.917	0.112	–	–	–
Lei et al. [31]	28.60	0.956	0.071	26.29	0.889	0.383
Ours (vanilla decoder + no cross-attn)	29.73	0.937	0.063	28.74	0.896	0.216
Ours (vanilla decoder + cross-attn)	30.23	0.940	0.062	29.84	0.918	0.187
Ours (cross-latent decoder + no cross-attn)	31.61	0.963	0.048	28.43	0.894	0.204
Ours (cross-latent decoder + cross-attn)	**32.06**	**0.964**	**0.042**	**30.37**	**0.937**	**0.161**

Table 1. **Quantitative Comparison With Competing Reflection Removal Methods**. This flash/no-flash dataset [31] contains ground truth images for the transmitted scene, captured by removing the glass. We compare our method against single image methods [9, 21, 69] and flash/no-flash-based methods [6, 31] over the metrics of separated transmission and reflection images. Our method achieves significantly better performance.

Figure 13. **Our Method Works Even Without Access to RAW Images.** Conventional Flash/no-flash methods [31, 61] require non-gamma-corrected RAW images as inputs to remove reflections. Likewise, our results shown previously are all using RAW images as input. However, some smartphones, e.g., some models of iPhone, do not give users access to RAW images, which limits the usage of flash/no-flash methods. In this experiment, we train our model to directly take in tonemapped flash/no-flash images as inputs, eliminating the need for RAW images. As shown here, our model still successfully performs reflection separation when applied to real-world tonemapped flash/no-flash image pairs, including the challenging Balcony scene (middle column) with reflections on a double-layer glass door.

5. Experimental Results

5.1. Experimental Setup

Our model is based on the Stable Diffusion architecture [43]. We used the pre-trained weights of Stable Diffusion v2 [43] (SD v2) to initialize both of our dual branch and cross-latent decoder. For dual branch training, we added our

Figure 14. **Inter-Branch Cross-Attention Enhances Reflection Separation.** Although the reflection estimated by the model without cross-attention (d) is much better than prediction from [31] (c), it still retains residuals from the transmitted scene (circled region). By contrast, the reflection with inter-branch cross-attention (e) closely resembles the ground truth (b).

Figure 15. **Our Cross-Latent Decoder Reduces hallucination**. Compared to the original pre-trained VAE encoder [26], our cross-latent decoder can leverage the high-frequency signal from the original captured image when it decodes the separated latent. As shown here, our cross-latent decoder's output preserves fine details faithful to the real scene, yielding clearer reconstructions, whereas the vanilla VAE decoder hallucinates the contents.

inter-branch cross-attention in the midblock attention of the UNet and follow the fine-tuning protocol of Marigold [23]. We used the same simulated and real datasets proposed in Lei et al. [31], which contains sets of flash/no-flash pairs and their corresponding ground truth transmission and reflection. We also evaluated the data from [61]. We first trained the dual branch model using a learning rate of 3×10^{-5}. We then trained our decoder with output images generated from our Stage 1 model and SD v2 decoder, using a learning rate of 10^{-5}. Both stages of our model were trained on a NVIDIA A6000 GPU, where Stage 1 took roughly two days and Stage 2 took one day. We provide training and inference details in Sec. 12 of the Suppl.

Compared Methods We conducted qualitative and quantitative comparisons using flash/no-flashed-based and pure software-based methods. We compared ours with the flash-based method [31], which consists of an optical flow network to handle misalignment and CNN networks for separation. We also compared with four recent software-based methods: [21], [9], [69] are single-image learning-based methods; while [8] is a burst imaging method based on neural rendering. Our results are presented in Fig 7, 8, 9, 10, with more results in Fig. 20,21, 22, 23 of the Suppl. We also compare our method quantitatively with compared methods in the real dataset in Lei et al. [31], achieving superior performance in all tested metrics (Tab. 1). We finally show an additional qualitative comparison with Lei et al. [31] and the ground truth transmission/reflection in Fig. 6.

6. Ablation Studies

Cross-Attention Tab. 1 and Fig. 14 show that the cross-attention module lets our model achieve separation performance (especially for the reflection component) both quantitatively and visually.

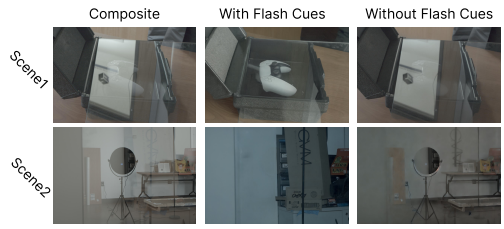

Figure 16. **Comparing Predicted Transmission of Flash/No-Flash vs. Single Image Models.** We train our model to only take a single composite image as the input. Compared to our flash/no-flash model, the single image model cannot effectively separate reflections. This illustrates the physical cues introduced from the flash/no-flash pair is crucial to our method's success.

Reflection Removal with Tonemapped Images Flash/no-flash reflection removal methods typically rely on RAW color space image inputs to preserve radiance intensity linearity. However, access to RAW images is limited on consumer devices (smartphones) without specialized software. We trained a variant of our model where RAW input is not needed, and instead taking the tonemapped flash/no-flash image as inputs. We show various real data evaluation results of our model in Fig. 13, where our model still works with tonemapped image inputs. A notable example is the Balcony scene, where the double-layer glass presents a challenging scenario due to multiple reflection paths.

Cross-Latent Decoder Fig. 15 shows that our cross-latent decoder enhances fine details within the separated images. While our Stage 1 latent separation model effectively isolates the transmission latent feature, the vanilla decoder induces hallucinations and blurriness, e.g., making text illegible as shown in Fig. 15. In contrast, our cross-latent decoder is able to extract high frequency details from the composite image based on the predicted latent from Stage 1's separation. We further analyze the roles of the two stages for separation in Sec. 9 of the Suppl.

Importance of Flash/No-Flash Fig. 16 shows that the flash cues are crucial for latent-space separation: we trained a variant of our model that takes in a single image as the input (instead of the flash/no-flash pair), and it failed to remove the reflections.

7. Conclusions

In conclusion, our Flash-Split method provides a robust solution for reflection separation in transparent surfaces, overcoming the need for precise flash/no-flash alignment. By performing reflection separation in the latent space, we effectively circumvent the flash/no-flash misalignment issue. We also employ a cross-latent decoding module to restore detailed and faithful features of the separated scenes from their latents. Evaluations on both simulated and challenging real-world data confirm our effectiveness, marking a substantial improvement in practical reflection separation.

Acknowledgements

This work was supported in part by a gift from Dolby Labs, AFOSR YIP award No. FA9550-22-1-0208, ONR award No. N000142312752, ARO ECP award No. W911NF2420113, and NSF CAREER award No. 2339616. We thank the reviewers for their useful feedback that helped improve this manuscript.

References

[1] Amit Agrawal, Ramesh Raskar, Shree K Nayar, and Yuanzhen Li. Removing photography artifacts using gradient projection and flash-exposure sampling. In *SIGGRAPH*, 2005. 1, 4

[2] Jean-Baptiste Alayrac, Joao Carreira, and Andrew Zisserman. The visual centrifuge: Model-free layered video representations. In *CVPR*, 2019. 3

[3] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Susstrunk. Single image reflection suppression. In *CVPR*, 2017. 1, 3

[4] Allison H. Baker, Alexander Pinard, and Dorit M. Hammerling. On a structural similarity index approach for floating-point data. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 11

[5] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. In *CVPR*, 2024. 5

[6] Yakun Chang, Cheolkon Jung, Jun Sun, and Fengqiao Wang. Siamese dense network for reflection 2021emoval with flash and no-flash image pairs. *IJCV*, 128, 2020. 7, 9

[7] Yakun Chang, Cheolkon Jung, Jun Sun, and Fengqiao Wang. Siamese dense network for reflection removal with flash and no-flash image pairs. *IJCV*, 128, 2020. 4, 9, 12, 13

[8] Ilya Chugunov, David Shustin, Ruyu Yan, Chenyang Lei, and Felix Heide. Neural spline fields for burst image fusion and layer separation. *CVPR*, 2024. 3, 6, 8

[9] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *ICCV*, 2021. 3, 6, 7, 8

[10] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*, 2017. 3

[11] Hany Farid and Edward H Adelson. Separating reflections and lighting using independent components analysis. In *CVPR*, 1999. 3

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 1981. 2, 3

[13] Yosef Gandelsman, Assaf Shocher, and Michal Irani. " double-dip": unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, 2019. 3

[14] Xiaojie Guo, Xiaochun Cao, and Yi Ma. Robust separation of reflection from multiple images. In *CVPR*, 2014.

[15] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *CVPR*, 2022. 3

[16] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 10

[18] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C Kot, and Boxin Shi. Panoramic image reflection removal. In *CVPR*, 2021. 3

[19] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C Kot, and Boxin Shi. Par2 net: End-to-end panoramic image reflection removal. *IEEE TPAMI*, 2023. 3

[20] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. In *NeurIPS*, 2021. 3

[21] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *ICCV*, 2023. 3, 6, 7, 8, 15

[22] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *ECCV*, 2024. 1, 5

[23] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 4, 5, 8, 10

[24] Eric Kee, Adam Pikielny, Kevin Blackburn-Matzen, and Marc Levoy. Removing reflections from raw photos. *ArXiv preprint abs/2404.14414*, 2024. 3

[25] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based rendering. *ArXiv preprint arXiv:1904.11934*, 2019. 3

[26] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, 2013. 2, 4, 5, 8

[27] Naejin Kong, Yu-Wing Tai, and Sung Yong Shin. High-quality reflection separation using polarized images. *IEEE TIP*, 20(12), 2011. 3

[28] Naejin Kong, Yu-Wing Tai, and Joseph S Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *IEEE TPAMI*, 36(2), 2013. 3

[29] Chenyang Lei and Qifeng Chen. Robust reflection removal with reflection-free flash-only cues. In *CVPR*, 2021. 1, 3

[30] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *CVPR*, 2020. 1, 3

[31] Chenyang Lei, Xudong Jiang, and Qifeng Chen. Robust reflection removal with flash-only cues in the wild. *IEEE TPAMI*, 45(12), 2023. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

[32] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE TPAMI*, 29(9), 2007. 3

[33] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. In *NeurIPS*, 2002. 1

[34] Anat Levin, Assaf Zomet, and Yair Weiss. Separating reflections from a single image using local features. In *CVPR*, 2004. 3

[35] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *CVPR*, 2020. 3

[36] Rui Li, Simeng Qiu, Guangming Zang, and Wolfgang Heidrich. Reflection separation via multi-bounce polarization state tracing. In *ECCV*, 2020. 3

[37] Yu Li and Michael S Brown. Exploiting reflection change for automatic reflection removal. In *ICCV*, 2013. 3

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

[39] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *CVPR*, 2020. 3

[40] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. In *NeurIPS*, 2019. 3

[41] Shree K Nayar, Xi-Sheng Fang, and Terrance Boult. Separation of reflection components using color and polarization. *IJCV*, 21(3), 1997. 3

[42] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *CVPR*, 2023. 4

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2021. 3, 4, 5, 7, 9, 10, 11, 12, 14, 15

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015. 4

[45] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *CVPR*, 2015. 1, 3

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 12, 13

[47] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 2, 3

[48] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Wen Gao, and Alex C Kot. Region-aware reflection removal with unified content and gradient priors. *IEEE TIP*, 2018. 3

[49] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *CVPR*, 2018.

[50] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, and Alex C Kot. Reflection scene separation from a single image. In *CVPR*, 2020.

[51] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, and Alex C Kot. Face image reflection removal. *IJCV*, 129, 2021. 3

[52] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. Stereodiffusion: Training-free stereo image generation using latent diffusion models. In *CVPR*, 2024. 5

[53] Tianfu Wang, Florian Schiffers, Florian Willomitzer, and Oliver Cossairt. A mitsuba-based study on trade-offs between projection and reflection based systems in structured-light 3d imaging. In *Computational Optical Sensing and Imaging*. Optica Publishing Group, 2021. 1

[54] Tianfu Wang, Menelaos Kanakis, Konrad Schindler, Luc Van Gool, and Anton Obukhov. Breathing new life into 3d assets with generative repainting. *ArXiv preprint arXiv:2309.08523*, 2023. 5

[55] Tao Wang, Wanglong Lu, Kaihao Zhang, Wenhan Luo, Tae-Kyun Kim, Tong Lu, Hongdong Li, and Ming-Hsuan Yang. Promptrr: Diffusion models as prompt generators for single image reflection removal. *ArXiv preprint arXiv:2402.02374*, 2024. 3

[56] Tianfu Wang, Jiazhang Wang, Nathan Matsuda, Oliver Cossairt, and Florian Willomitzer. Differentiable deflectometric eye tracking. *IEEE TCI*, 10, 2024. 1

[57] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *CVPR*, 2019. 3

[58] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *CVPR*, 2019. 3

[59] Zhihao Xia, Michaël Gharbi, Federico Perazzi, Kalyan Sunkavalli, and Ayan Chakrabarti. Deep denoising of flash and no-flash pairs for photography in low-light environments. *CVPR*, 2020. 3

[60] Zhihao Xia, Jason Lawrence, and Supreeth Achar. A dark flash normal camera. *ICCV*, 2021. 3

[61] Mingyang Xie, Haoming Cai, Sachin Shah, Yiran Xu, Brandon Y. Feng, Jia-Bin Huang, and Christopher A. Metzler. Flash-splat: 3d reflection removal with flash cues and gaussian splats. In *ECCV*, 2024. 1, 3, 7, 8, 13, 14

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

[63] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *ECCV*, 2018. 3

[64] Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, and Weiyu Xu. Fast single image reflection suppression via convex optimization. In *CVPR*, 2019. 1, 3

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 6

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 11

[67] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018. 3

[68] Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, and Alex C Kot. Single image reflection removal with absorption effect. In *CVPR*, 2021.

[69] Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, and Bo Li. Revisiting single image reflection removal in the wild. In *CVPR*, 2024. 1, 6, 7, 8

[70] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. In *CVPR*, 2020. 3