FIRM: Flexible Interactive Reflection ReMoval

Xiao Chen[1,3], Xudong Jiang[2], Yunkang Tao[3],
Zhen Lei[3,4,5], Qing Li[1], Chenyang Lei[3]*, Zhaoxiang Zhang[3,4,5]*
[1]The Hong Kong Polytechnic University [2]ETH Zurich
[3]Center for Artificial Intelligence and Robotics, HKISI-CAS
[4] Institute of Automation, Chinese Academy of Sciences
[5] University of Chinese Academy of Sciences

Abstract

Removing reflection from a single image is challenging due to the absence of general reflection priors. Although existing methods incorporate extensive user guidance for satisfactory performance, they often lack the flexibility to adapt user guidance in different modalities, and dense user interactions further limit their practicality. To alleviate these problems, this paper presents **FIRM**, a novel framework for **F**lexible **I**nteractive image **R**eflection re**M**oval with various forms of guidance, where users can provide sparse visual guidance (e.g., points, boxes, or strokes) or text descriptions for better reflection removal. Firstly, we design a novel user guidance conversion module (UGC) to transform different forms of guidance into unified contrastive masks. The contrastive masks provide explicit cues for identifying reflection and transmission layers in blended images. Secondly, we devise a contrastive mask-guided reflection removal network that comprises a newly proposed contrastive guidance interaction block (CGIB). This block leverages a unique cross-attention mechanism that merges contrastive masks with image features, allowing for precise layer separation. The proposed framework requires only 10% of the guidance time needed by previous interactive methods, which makes a step-change in flexibility. Extensive results on public real-world reflection removal datasets validate that our method demonstrates state-of-the-art reflection removal performance.

Introduction

Image reflection removal refers to the task of eliminating unwanted reflections in images captured through glass. Specifically, the partially reflective glass superposes the scene of interest with reflections behind the observer, which reduces image contrast and potentially obscures important details. Extensive research on image reflection removal primarily focuses on low-level and physics-based priors, such as gradient sparsity (Levin and Weiss 2007), ghosting effect (where duplicate elements appear on thick glasses) (Shih et al. 2015), and reflection blurriness (Fan et al. 2017; Yang et al. 2019). However, these methods often struggle with *beyond-assumption* reflections (e.g., sharp reflections), due to the similarity in natural image statistics between transmission and reflection layers.

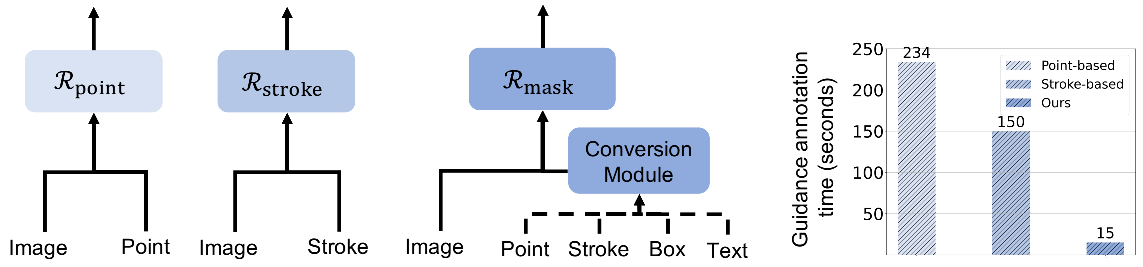
To alleviate the inherent ambiguity in layer separation, using auxiliary inputs as additional guidance has become a trend. Several works utilize multiple images or sensors to gather additional information about reflections, such as polarization images (Patrick et al. 2018; Lei et al. 2020; Kong, Tai, and Shin 2014; Lyu et al. 2019; Rui et al. 2020), flash images (Lei and Chen 2021), and multi-view images (Xue et al. 2015; Niklaus et al. 2021; Han and Sim 2017). However, these methods require additional sensors or multiple captures, limiting their flexible applications in practice.

Interactive methods (Levin and Weiss 2007; Zhang et al. 2020) have also been studied, enabling reflection removal with more readily available human guidance, yet they exhibit significant limitations: i) support only a specific form of user guidance, and ii) require dense interactions for satisfactory performance, leading to high time costs. For instance, in (Zhang et al. 2020), users draw dense strokes on the edge of reflection and background, resulting in nearly 150 seconds of time cost per image, as indicated in Figure 1.

To address these limitations, we present FIRM, a novel interactive framework that supports flexible user guidance forms, including point, stroke, box, and text, for guiding reflection removal. As shown in Figure 1, unlike previous interactive methods, our reflection removal network is not limited to specific guidance forms, as it incorporates a conversion module to unify various guidance into a mask format. Moreover, users can specify reflection and transmission layers with sparse guidance in an average of 15 seconds per image, significantly reducing the time cost from 234 seconds required by prior methods (Levin and Weiss 2007).

Specifically, we propose a two-stage pipeline in FIRM. **Firstly**, we propose the **user guidance conversion** (UGC) module to convert different guidance into a unified format, that is, segmentation mask. For text guidance, we adopt the text-based segmentation model (Lai et al. 2023). For visual guidance (i.e., point, box, stroke), we develop a novel Segment Any Reflection Model (SARM) based on the Segment Anything Model (SAM) (Kirillov et al. 2023), which freezes most parameters of SAM and updates only a learnable token and a feature selection block in the mask decoder. We do this because we observe that the original SAM falters in blended images when provided with sparse point prompts, as shown

[*]Corresponding authors: Chenyang Lei (leichenyang7@gmail.com), Zhaoxiang Zhang (zhaoxiang.zhang@ia.ac.cn)


$\mathcal{R}_{\text{point}}$
$\mathcal{R}_{\text{stroke}}$
$\mathcal{R}_{\text{mask}}$
Conversion Module
Image
Point
Image
Stroke
Image
Point
Stroke
Box
Text
Guidance annotation time (seconds)
250
200
150
100
50
234
150
15
Point-based
Stroke-based
Ours

(a) Guidance-specific interactive reflection removal

(b) Flexible interactive reflection removal **(Ours)**

(c) Comparison of guidance annotation time

Figure 1: **Comparison between previous interactive method (Levin and Weiss 2007; Zhang et al. 2020) and ours. (a)** and **(b)** illustrate the structural differences. The previous methods are guidance-specific, with tailored reflection removal networks ($\mathcal{R}_{\text{point}}$, $\mathcal{R}_{\text{stroke}}$) for each guidance form(e.g., point or stroke). In contrast, our framework is flexible, utilizing a conversion module to accommodate various forms of guidance by transforming them into a unified "segmentation mask". **(c)** Additionally, we compare the time cost of providing user guidance, where our method requires significantly less time per image than the results reported in previous works (Zhang et al. 2020).

in Table 3. To address the performance degradation of SAM on blended images while maintaining its strong zero-shot capability, our SARM is trained using a lightweight parameter tuning strategy. Once trained, by prompting the UGC module with guidance on both transmission and reflection regions, we can obtain corresponding masks, which together form contrastive masks. **Secondly**, we design a **contrastive mask-guided reflection removal network** that employs a novel contrastive guidance interaction block. This block enables the contrastive mask to interact with blended features and precisely separate out transmission and reflection features using cross-attention mechanisms.

To evaluate the efficacy of our proposed FIRM framework, we augment established benchmark datasets (Zhang et al. 2018; Wan et al. 2017) by incorporating additional user guidance, contributing to the first comprehensive interactive reflection removal dataset. Empirical results confirm that FIRM effectively improves reflection removal performance while requiring significantly less human guidance. The main contributions of this work are threefold:

- We propose the first universal framework FIRM for interactive image reflection removal, supporting diverse flexible forms of guidance. In particular, we develop the UGC module with a tailored segmentation model SARM, which enhances the ability to generate accurate reflection masks with sparse visual guidance.
- We propose a novel reflection removal network that uses contrastive masks as additional guidance, employing a cross-attention mechanism to fuse transmission and reflection masks with blended image features. Extensive experiments demonstrate that it achieves superior reflection removal performance while requiring 10 × less time for annotating user guidance.
- We contribute a comprehensive benchmark dataset for interactive image reflection removal, consisting of four forms of raw user guidance and their converted segmentation masks, facilitating further study in this field.

Related Work

Single-image reflection removal. Single-image reflection removal is challenging due to its ill-posed nature, which often leads to ambiguous decompositions, as explored in (Wan et al. 2017, 2022). Traditional methods rely on defocused and ghosting cues. The defocus cue refers to reflections appearing blurry when focusing on the transmission layer due to depth disparity. Non-learning based methods (Yang et al. 2019) exploit this by suppressing reflections with image gradient statistics, while learning-based methods like (Fan et al. 2017; Zhang et al. 2018) use these assumptions for data synthesis. The ghosting cue (Shih et al. 2015) is relevant for thick glass, identifies multiple reflections on the glass surface. However, these methods face limitations when these assumptions fail. Though several approaches employ GANs (Wen et al. 2019; Ma et al. 2019; Goodfellow et al. 2014) or more accurate physical rendering methods (Kim, Huo, and Yoon 2020) to mimic real reflection distributions, or directly collect real-world data (Zhang et al. 2018; Wei et al. 2019a; Li et al. 2020; Lei et al. 2021), they still face challenges in covering diverse kinds of reflections (Lei et al. 2020; Hu and Guo 2023; Zhu et al. 2024), underscoring the need for further research.

Reflection removal with auxiliary inputs. Alternative methods that use additional inputs have been explored. Motion-based techniques leverage multiple images to capture distinct motion characteristics, which aids in separating reflections. However, they require complex image capture setups and are limited by specific assumptions (Guo, Cao, and Ma 2014; Han and Sim 2017; Li and Brown 2013; Liu et al. 2020; Sun et al. 2016; Xue et al. 2015; Niklaus et al. 2021; Chugunov et al. 2023). Polarization-based methods leverage different polarization properties of reflection and transmission (Farid and Adelson 1999; Kong, Tai, and Shin 2014; Patrick et al. 2018; Lyu et al. 2019; Li et al. 2020; Rui et al. 2020). Flash/ambient image pairs have also been studied to handle reflections and shadows (Agrawal et al. 2005; Chang et al. 2020; Lei, Jiang, and Chen 2023). These meth-

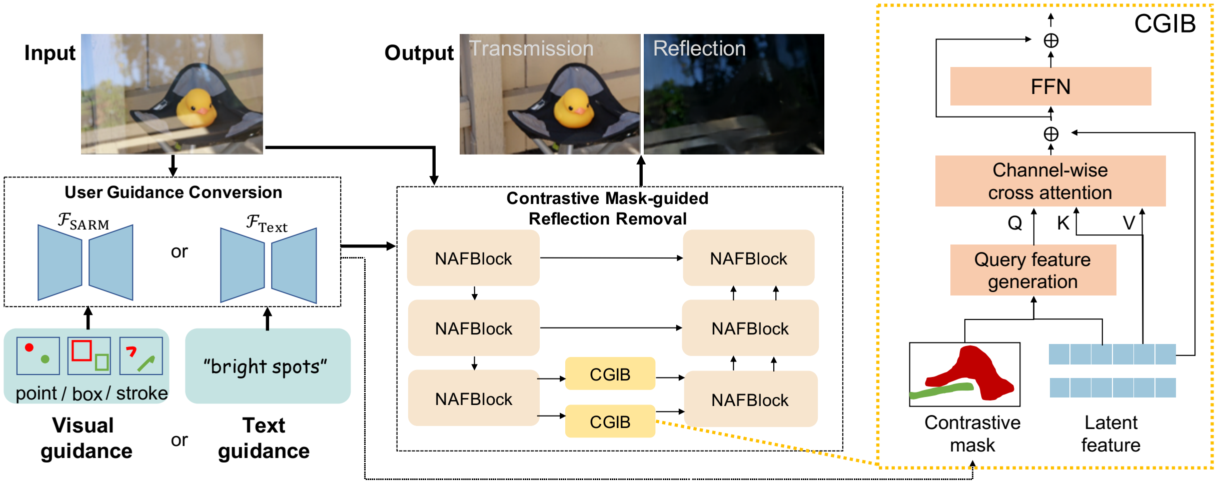


Figure 2: **Illustration of our proposed pipeline FIRM.** FIRM receives a blended image with diverse forms of user guidance, such as visual guidance or text descriptions. The user guidance conversion module (**UGC**) first transforms the raw input into contrastive masks with the user guidance. Then, the contrastive mask-guided network, incorporated with our designed Contrastive Guidance Interaction Block (**CGIB**) blocks, utilizes contrastive masks to separate the transmission and reflection layers from the blended input. (Detailed network configurations are provided in supplementary materials.)

ods generally require additional equipment or specific image acquisition conditions. Interactive methods utilize user guidance, such as dense strokes or points, as additional input, but they require extensive user annotations (Zhang et al. 2020; Levin and Weiss 2007; Chen et al. 2024b). In most recent work (Zhong et al. 2024), text descriptions are introduced as a form of high-level user guidance. Our work diverges by offering a unified user guidance representation that accommodates various guidance forms, enhancing flexibility in user interactions.

Method

Overview

In Figure 1, we present the limited practicality of previous interactive methods, which motivates us to design two key objectives for the new framework: First, user guidance should be flexible and support various forms; Second, it ensures fast and convenient guidance annotation to further improve the interactive process. As illustrated in Figure 2, given the blended image and the raw user guidance in flexible modalities, the proposed FIRM framework predicts the underlying reflection and transmission images in two stages. First, the user guidance conversion module transforms the inputs into a unified contrastive mask, which captures prominent reflection and transmission region information. Then, the contrastive mask-guided transformer, built upon the Contrastive Guidance Interactive Block (CGIB) as its core component, integrates image features with the contrastive masks for precise decomposition.

UGC: User Guidance Conversion

To enhance the flexibility of utilizing various forms of user guidance, we introduce the UGC module to transform the user guidance $g^i \in \mathbf{G}$ into a unified mask format, where $\mathbf{G} = \{g^i\}_{i=1}^{N}$ is the set of N user guidance (i.e., point, box, stroke, text). The UGC module consists of interactive segmentation models, represented as $\mathcal{F} = \{\mathcal{F}_{\text{SARM}}, \mathcal{F}_{\text{Text}}\}$, where we propose a novel Segment Any Reflection Model (SARM) $\mathcal{F}_{\text{SARM}}$ for handling visual guidance and an off-the-shelf segmentation model $\mathcal{F}_{\text{Text}}$ for text guidance (Lai et al. 2023). The workflow can be formulated as:

$$\mathbf{M} = \mathcal{F}(\mathbb{S}(g^i), \mathbf{I}), \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ is the blended image and power set $\mathbb{S}(g^i)$ represents the corresponding mix set of user guidance, and segmentation mask $\mathbf{M} \in \{0, 0.5, 1\}^{H\times W\times 1}$ contains distinct values for reflection (1), transmission (0.5) and non-annotated (0) areas. Specifically, we propose SARM to address blended images while preserving the strong zero-shot learning capabilities of SAM (Kirillov et al. 2023).

Preliminary of SAM. In the original SAM, the image encoder uses the Vision Transformer to process input images, and the prompt encoder handles sparse prompts (e.g., points, boxes) by converting them into suitable latent representations. The mask decoder then combines image and prompt embeddings with an output token using a two-way transformer module. It then applies transpose convolutions to upsample mask features and utilizes token-to-image attention to generate an output token for each mask. Finally, an MLP converts this output token into a dynamic classifier, which is multiplied with the mask features to produce the final segmentation mask.

The tailored SARM. To achieve more accurate segmentation on blended images using sparse visual prompts, we tailored SARM with minimal additional parameters to SAM, keeping SAM's image and prompt encoders fixed to maintain zero-shot capabilities while making two key modifications in the mask decoder. First, we introduce the learnable degradation-invariant token. This token (size of 1×256) is

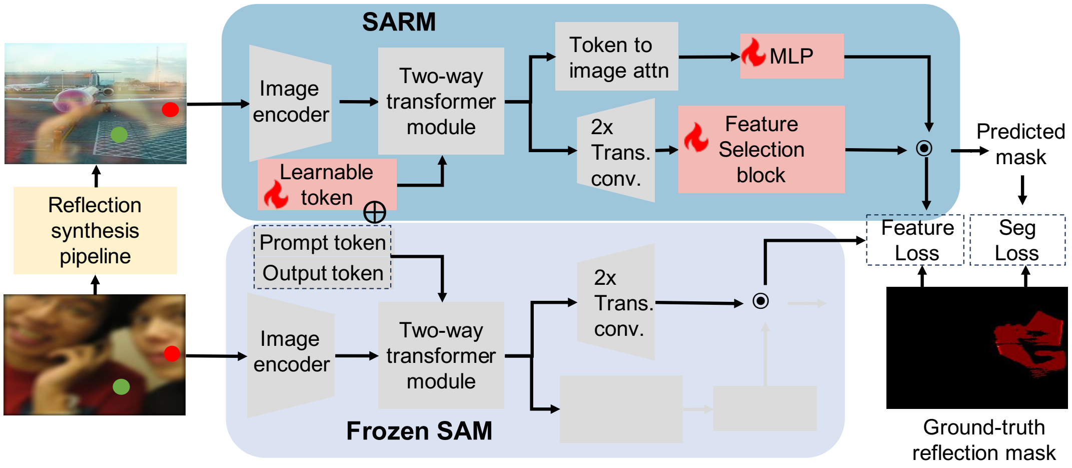


Figure 3: **Illustration of the training pipeline of SARM.** We introduce learnable *degradation-invariant token* and *feature selection block* into the original SAM architecture, aiming for accurate mask prediction in blended images. To maintain the zero-shot capability of SAM (Kirillov et al. 2023), only a limited number of parameters in the mask decoder are trainable, while the parameters of the image encoder and prompt encoder from the pre-trained SAM remain fixed.

concatenated with SAM's output tokens (size of 4×256) and prompt tokens (size of $N_{\text{prompt}} \times 256$) as the input to mask decoder. Additionally, we incorporate a learnable three-layer MLP to generate dynamic weights, which are then used in a point-wise product with the mask features. Second, we design the feature selection block to enhance prominent reflection features in blended images. This block takes intermediate mask features $\mathbf{F} \in \mathbb{R}^{c \times h \times w}$, first squeezing the spatial information into $\mathbf{F_{avg}} \in \mathbb{R}^{c \times 1 \times 1}$ via average pooling. It then employs a lightweight gating mechanism with sigmoid activation function $\sigma(.)$, as follows:

$$\tilde{\mathbf{F}} = \sigma(\mathbf{W_1}(\text{GELU}(\mathbf{W_0}(\mathbf{F_{avg}}))) \odot \mathbf{F}, \qquad (2)$$

where $\mathbf{W_0} \in \mathbb{R}^{c/r \times c}$, $\mathbf{W_1} \in \mathbb{R}^{c \times c/r}$ denote the learnable MLP weights. By exploiting the non-linear inter-channel relationship of mask features, the network learns to focus on channels that resemble salient features.

The training pipeline is shown in Figure 3. We first select one clear image and input it into SAM. This image is also used to synthesize the blended image, which is then fed into SARM. The proposed SARM is supervised by both feature-level and mask-level loss. For the mask-level segmentation loss, we follow the configuration in (Kirillov et al. 2023), combining Dice $\mathcal{L}_{\text{Dice}}$ (Sudre et al. 2017) and Focal Loss $\mathcal{L}_{\text{Focal}}$ (Lin et al. 2017). Additionally, we design a mask feature consistency loss to enhance the extraction of prominent reflection features. The overall loss function is as follows:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}}(\mathbf{M_p}, \mathbf{M_{gt}}) + \lambda_0 \mathcal{L}_{\text{Focal}}(\mathbf{M_p}, \mathbf{M_{gt}}) \qquad (3)\\ + \lambda_1 \mathcal{L}_{\text{MSE}}(\tilde{\mathbf{F}}_p \odot \mathbf{M_{gt}}, \tilde{\mathbf{F}_{gt}} \odot \mathbf{M_{gt}}),\end{aligned}$$

where $\mathbf{M_p}$, $\mathbf{M_{gt}}$ denotes the predicted and ground-truth reflection mask, $\tilde{\mathbf{F}}_\mathbf{p}$ and $\tilde{\mathbf{F}_\mathbf{gt}}$ represent the mask features of SAM and SARM, λ_0 and λ_1 are hyper-parameters for different loss terms.

Algorithm 1: Training data synthesis pipeline

```
Input: Two clear RGB images $\mathbf{T}$, $\mathbf{R}$ and its instance mask $\mathbf{M}$
Output: Blended image $\mathbf{I}$, reflection instance mask $\mathbf{M_r}$, contrastive points $\{p^{\text{pos}}, p^{\text{neg}}\}$
$\mathbf{I} \leftarrow$ Reflection_Synthesis($\mathbf{T}, \mathbf{R}$)
$\mathbf{R}' \leftarrow$ threshold($\mathbf{I} - \mathbf{T}, 0$) # Residual map
max_reflection_value $\leftarrow 0$
for each instance $\mathbf{M_i}$ in $\mathbf{M}$ do
    avg_value $\leftarrow$ MEAN($\mathbf{R}' \cdot \mathbf{M_i}$)
    if avg_value > max_reflection_value then
        max_reflection_value $\leftarrow$ avg_value
        $\mathbf{M_r} \leftarrow \mathbf{M_i} \cdot \mathbf{R}'$
    end if
end for
Randomly sample a reflection point $p^{pos}$ from $\mathbf{M_r}$
Randomly select a transmission point $p^{neg}$ from the neighbour of $\mathbf{M_r}$
return $\mathbf{I}$, $\mathbf{M_r}$, $\{p^{\text{pos}}, p^{\text{neg}}\}$
```

Constructing training data for SARM. Since there is no public reflection segmentation dataset, we manually synthesize training data based on the COCO dataset (Lin et al. 2014). As illustrated in Algorithm 1, we first apply the pipeline from (Zhang et al. 2018) to synthesize blended images using two clear images. Then we obtain a pseudo reflection instance mask $\mathbf{M_r}$ by traversing the instance mask with the highest values in the residual map, along with contrastive points located inside and outside the reflection mask. The proposed SARM is trained with blended images using contrastive points as prompts.

Inference. SARM supports points, boxes, or strokes as prompts. When the reflection area is labeled positively, we obtain the reflection mask; otherwise, we get the transmis-

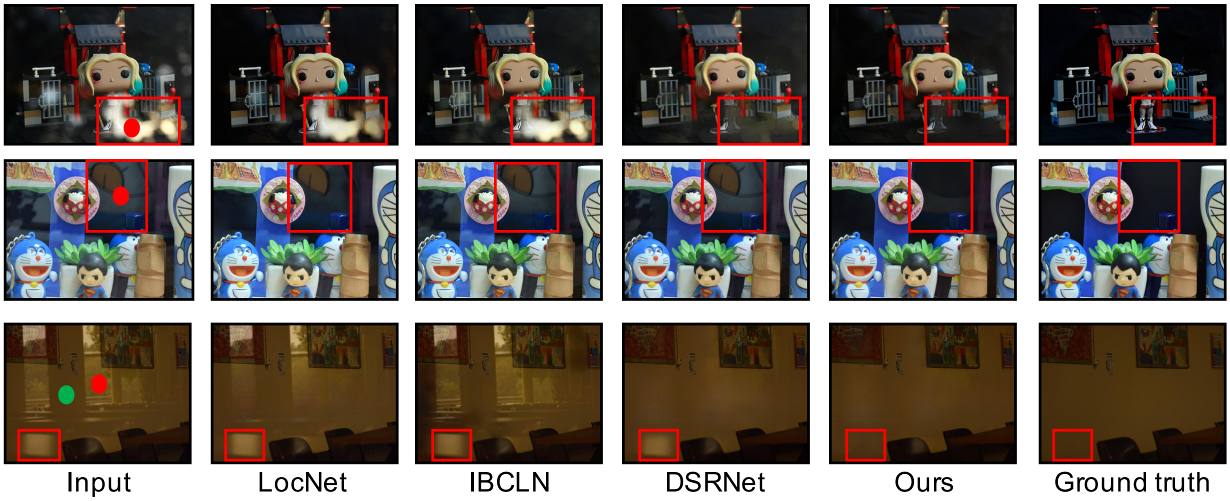


Figure 4: **Qualitative comparison of estimated transmissions between representative single-image-based methods and ours on Real20 and SIR2 datasets.** Single-image-based methods struggle to remove sharp reflections. Our approach achieves much better reflection removal than baselines with very sparse point guidance on reflection and transmisson areas.

sion mask. Stroke guidance is supported by uniformly sampling points along the stroke trajectory. Finally, we merge the reflection and transmission masks into a single mask, referred to as the contrastive mask.

Contrastive Mask-Guided Reflection Removal

In the second stage, building upon a U-shaped encoder-decoder architecture, we propose a novel Contrastive Guidance Interaction Block (CGIB) that effectively incorporates guidance information from contrastive masks into the feature decomposition process. Our method differs from existing mask-guided methods (Dong et al. 2021b) by integrating image features with auxiliary contrastive masks that provide more accurate region boundary information, enabling more precise layer separation.

Specifically, with the blended image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and the converted contrastive mask $\mathbf{M} \in \mathbb{R}^{H\times W\times 1}$, we concatenate them along the channel dimension and feed it into the reflection removal network $\mathcal{R}$, which separates transmission layer $\hat{\mathbf{T}}$ and reflection layer $\hat{\mathbf{R}}$. The network primarily consists of NAFNet Blocks (Chen et al. 2022) for feature extraction, incorporating CGIB into the middle layers for facilitating feature decoding. The overall process is:

$$\{\hat{\mathbf{T}}, \hat{\mathbf{R}}\} = \mathcal{R}(\mathbf{I} \oplus \mathbf{M}, \mathbf{M}; \theta), \tag{4}$$

where θ denotes the network parameters of $\mathcal{R}$ and $\oplus$ denotes the concatenation operation.

The CGIB consists of three components, namely, Query Feature Generation, Channel-wise Cross Attention (CCA), and Feed-Forward Network (FFN). First, a query is generated using the contrastive mask. The contrastive mask is resized into $\mathbf{M}^{'} \in \mathbb{R}^{h\times w\times 1}$ to match the spatial dimensions of the blended features, followed by element-wise multiplication between the resized mask and blended features $\mathbf{I}_\theta$. This step extracts the prominent reflection or transmission feature as queries, denoted as $\mathbf{Q}_\theta$. Note that, to handle the different resolutions of input images during inference, we resize $\mathbf{Q}_\theta$ to a fixed spatial dimension $h' \times w'$, Next, the CCA module uses the query feature $\mathbf{Q}_\theta$ as anchors to correlate similar components in the blended features $\mathbf{I}_\theta$. The overall process of the CCA module is:

$$\mathrm{CCA}(\mathbf{Q}_\theta, \mathbf{K}, \mathbf{V}) = \mathbf{V}\,\mathrm{Softmax}(\frac{\mathbf{Q}_\theta \mathbf{K}^\top}{\alpha}), \tag{5}$$

where $\mathbf{K} \in \mathbb{R}^{c\times h'w'}$ and $\mathbf{V} \in \mathbb{R}^{hw\times c}$ denote the blended image feature-generated key and value projections respectively, α is a temperature factor. The FFN component design strictly follows previous work (Zamir et al. 2022). The whole network is trained with pixel-wise reconstruction loss in the image and gradient domain (Hu and Guo 2023), perceptual loss (Wei et al. 2019b), and exclusion loss (Zhang et al. 2018).

Experiments

Dataset

Following the setting in (Hu and Guo 2023), the training data for reflection removal consists of 7,643 synthesized pairs from the PASCAL VOC dataset (Everingham et al. 2010) and 90 real pairs from (Zhang et al. 2018). The proposed FIRM is trained using point guidance. For real data, we manually label one reflection and one transmission point per image. For synthetic data, we obtain contrastive points following the pipeline in Algorithm 1. The test data includes **Real20** and **SIR2** (Zhang et al. 2018; Wan et al. 2017). The *SIR2* dataset (Wan et al. 2017) consists of three data splits: *SIR2-Object*, *SIR2-Postcard*, and *SIR2-Wild*, each featuring distinct contents and depth scales.

Flexible interactive reflection removal dataset. Since there is no publicly available evaluation dataset for interactive image reflection removal, we construct a comprehensive dataset that includes four forms of guidance. This

Category	Method	Real20 (20)		Object (200)		Postcard (199)		Wild (55)		Average	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Single-image-based methods	Zhang et al. (Zhang et al. 2018)	22.55	0.788	22.68	0.879	16.81	0.797	21.52	0.832	20.08	0.835
	BDN (Yang et al. 2018)	18.41	0.726	22.72	0.856	20.71	0.859	22.36	0.830	21.65	0.849
	ERRNet (Wei et al. 2019a)	22.89	0.803	24.87	0.896	22.04	0.876	24.25	0.853	23.53	0.879
	IBCLN (Li et al. 2020)	21.86	0.762	24.87	0.893	23.39	0.875	24.71	0.886	24.10	0.879
	RAGNet (Li et al. 2023)	22.95	0.793	26.15	0.903	23.67	0.879	25.53	0.880	24.90	0.886
	DMGN (Feng et al. 2021)	20.71	0.770	24.98	0.899	22.92	0.877	23.81	0.835	23.80	0.877
	Zheng et al. (Zheng et al. 2021)	20.17	0.755	25.20	0.880	23.26	0.905	25.39	0.878	24.19	0.885
	YTMT (Hu and Guo 2021)	23.26	0.806	24.87	0.896	22.91	0.884	25.48	0.890	24.05	0.886
	LocNet (Dong et al. 2021a)	23.34	0.812	24.36	0.898	23.72	0.903	25.73	0.902	24.18	0.893
	DSRNet (Hu and Guo 2023)	23.85	0.813	26.38	0.918	24.56	0.908	24.79	0.896	25.46	0.908
Interactive methods	Levin et al.† (Levin and Weiss 2007)	-	-	-	-	-	-	-	-	-	0.798
	FGNet (Zhang et al. 2020)	24.21	0.812	24.97	0.876	22.84	0.863	25.23	0.891	24.07	0.872
	Zhong et al. (Zhong et al. 2024)	24.05	0.814	26.41	0.920	24.62	0.905	26.20	0.920	25.53	0.909
	Ours	**26.88**	**0.826**	**26.80**	**0.921**	**24.90**	**0.910**	**29.71**	**0.932**	**26.34**	**0.914**

Table 1: **Quantitative comparison with baselines on Real20 and SIR2 dataset.** Our method trained with points (i.e., Ours-point) achieves the best performance on most evaluated datasets. We notice the reflection images in *SIR2-Postcard* tend to be more blurry, which makes the performance difference smaller. In wild scenes like *Real20*, *SIR2-Object*, and *SIR2-Wild*, where the reflections are sharper, the improvement of our approach is larger. For (Levin and Weiss 2007), we reference the results from (Zhang et al. 2020), indicated by †.

dataset builds on the public reflection datasets Real20 and SIR2 (Zhang et al. 2018; Wan et al. 2017), where we further annotate prominent reflection and transmission areas in the blended images. We engage a team of annotators to label points, strokes, and bounding boxes on blended images. We then extract point coordinates from these annotations and feed them into the trained SARM as prompts to obtain corresponding segmentation masks. Text-guided segmentation masks are generated using the model (Lai et al. 2023), with text descriptions manually labeled by the annotators.

Implementation Details

The proposed framework is implemented with PyTorch. During the training phase of SARM, only the proposed modules are optimized. Using point-based prompts, SARM is trained with a fixed learning rate of 0.0005 for 50 epochs on 8 NVIDIA A100 GPUs. The batch size is set as 8. The reflection removal network is optimized using the Adam optimizer for a total of 200,000 iterations, with a batch size of 8 on a single A100 GPU. The initial learning rate is set to 10^{-3} and gradually reduce to 10^{-6} with the cosine annealing schedule (Loshchilov and Hutter 2016).

Evaluations on Reflection Removal

We first compare the reflection removal performance of the proposed FIRM with two categories of methods, including single-image-based and interactive methods.

Baselines. i) *Single-image-based methods*, including Zhang et al. (Zhang et al. 2018), BDN (Yang et al. 2018), ERRNet (Wei et al. 2019a), IBCLN (Li et al. 2020), RAGNet (Li et al. 2023), DMGN (Feng et al. 2021), Zheng et al. (Zheng et al. 2021), YTMT (Hu and Guo 2021), LocNet (Dong et al. 2021b), DSRNet (Hu and Guo 2023). **ii)** *Interactive methods* (Levin and Weiss 2007; Zhang et al. 2020; Zhong et al. 2024). We retrain these methods on our training data for fair comparisons if their codes are available.

Quantitative results. In Table 1, we present the quantitative results of our approach and baselines on *Real20* (Zhang et al. 2018) and *SIR2* (Wan et al. 2017) dataset. We employ PSNR (Huynh-Thu and Ghanbari 2008) and SSIM (Wang, Simoncelli, and Bovik 2003) as metrics to evaluate the recovery quality of transmission layers. Our proposed FIRM (with points as guidance), denoted as "Ours," consistently outperforms other methods in all data sets, showcasing its superior generalization ability and effectiveness. Notably, our method also surpasses the text-based interactive approach (Zhong et al. 2024). We speculate this arises from the recognizable layer ambiguity, where certain image layers lack corresponding language descriptions. In contrast, point annotations provide greater flexibility across diverse scenarios. Additionally, our approach significantly reduces the reliance on dense point annotations, reducing the annotation number from 50 (Levin and Weiss 2007) points per image to only 2 points. This sparse point guidance not only simplifies the user interaction process but also enhances the practicality of our method in real-world applications.

Qualitative results. We provide qualitative comparisons with single-image-based methods in Figure 4. As depicted, single-image methods often struggle to separate sharp reflections from the input image. For instance, bright spots on the cartoon dolls (row 1) and walls (row 3), and the pillow with orange patterns (row 2) have similar intensities as foregrounds. In contrast, our method is capable of producing high-quality transmission images. We also present results from interactive methods in Figure 5, including FGNet (Zhang et al. 2020), which requires dense scribbles and yields inferior results, and the method (Zhong et al. 2024), which uses text descriptions for reflections and transmission as additional guidance. For reflections lacking describable semantics (indicated as "Not provided"), the text-based method struggles to identify them, while our approach uses just two or three sparse points to achieve significantly better visual quality.

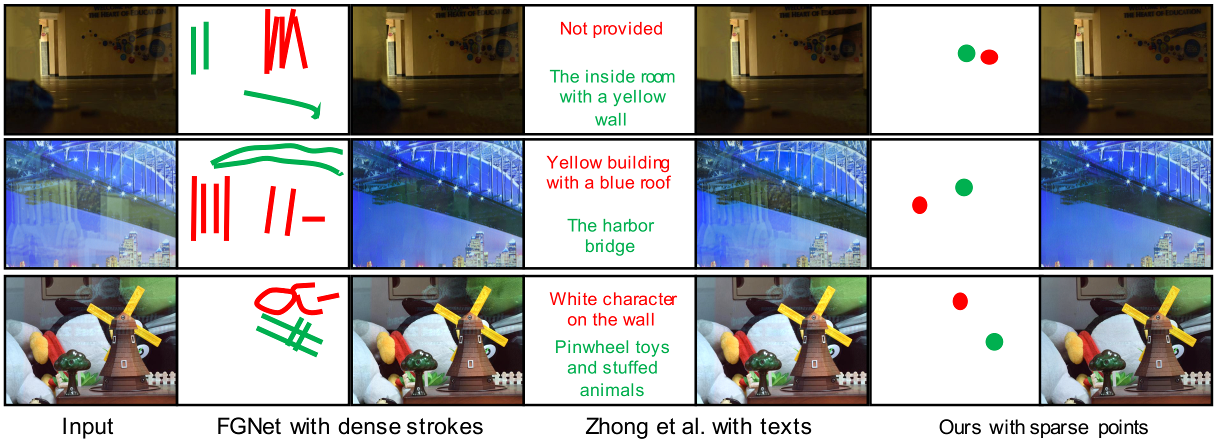

Not provided
The inside room with a yellow wall
Yellow building with a blue roof
The harbor bridge
White character on the wall
Pinwheel toys and stuffed animals
Input
FGNet with dense strokes
Zhong et al. with texts
Ours with sparse points

Figure 5: **Qualitative comparison of predicted transmissions between state-of-the-arts interactive methods and ours on SIR2 datasets (Wan et al. 2017).** The guidance for reflection and transmission regions is labeled with different colors. Our approach achieves superior reflection removal using just 2 sparse points.

Method	NAFNet	UGC	CGIB	Contrastive Mask	PSNR	SSIM
Blended Only	✓				24.01	0.880
Raw Point	✓				24.05	0.886
Raw Mask	✓	✓			25.70	0.907
Reflection Mask	✓	✓	✓		26.07	0.902
Ours	✓	✓	✓	✓	**26.34**	**0.914**

Table 2: **Ablation study of the proposed FIRM.** Ablation results show that raw prompt-based methods underperform, while feature-level interactions with converted masks achieve better results across most datasets.

Ablation Study

We conduct ablation studies to validate the effectiveness of our proposed modules or designs, including **UGC**, **CGIB**, and the **Contrastive Mask**. All model variants are trained from scratch using the same NAFNET-based architecture (Chen et al. 2022) and evaluated on the Real20 and SIR2 datasets using points as additional guidance. These method variants include: *i) Blended Only*: Using only blended images as input to the network; *ii) Raw Point*: Without UGC and CGIB, raw points are directly combined with blended images; *iii) Raw Mask*: Without CGIB, the converted contrastive masks are directly combined with blended images as input; *iv) Reflection Mask*: Only the converted reflection mask is used for deep feature interaction in CGIB. The average performance is shown in Table 2. Integrating raw points with the blended image directly or using blended image only yields inferior results, indicating the converted segmentation mask (UGC) is more effective for guiding removal. Directly combining converted masks with blended images also results in limited gains, emphasizing the importance of deep feature interaction (CGIB). Further, using reflection masks only for feature interaction cannot achieve optimal performance due to the lack of interaction cues.

We also evaluate the segmentation performance of the trained SARM on synthesized reflections using the COCO validation set (Lin et al. 2014) and real-world reflections from SIR2 (Wan et al. 2017). For comparison, we include RobustSAM (Chen et al. 2024a), a recent model designed for degraded image segmentation. Unlike common degradations, reflections usually exhibit arbitrary patterns, which makes our proposed SARM more suitable for reflection segmentation, as shown in Table 3.

Method	Synthetic Data				SIR2			
	Reflection		Transmission		Reflection		Transmission	
	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice
Frozen-SAM	0.306	0.469	0.840	0.905	0.326	0.432	0.803	0.879
RobustSAM	0.316	0.472	0.844	0.912	0.337	0.454	0.817	0.894
SAM-decoder-ft	0.369	0.505	0.854	0.923	0.449	0.574	0.824	0.903
SARM	**0.429**	**0.565**	**0.857**	**0.926**	**0.541**	**0.702**	**0.829**	**0.906**

Table 3: **Segmentation comparison on the synthetic data based on COCO validation set and real data on SIR2-dataset using point prompts.** "-decoder-ft": finetuning the entire SAM mask decoder.

Conclusion

This paper proposes a flexible interactive reflection removal approach that leverages human guidance in diverse forms as an auxiliary input. The user guidance conversion module, built upon a novel segment-any-reflection model, generates accurate reflection masks while preserving strong performance on clear images. Further, a Contrastive Guidance Interaction Block is designed in an encoder-decoder-based network to facilitate precise image layer separation using the generated masks, achieving superior reflection removal performance across various datasets. This highlights the significance of human guidance in addressing ambiguity in single-image reflection removal. Furthermore, we enhance existing public reflection removal datasets with sparse human annotations, facilitating further study.

Acknowledgments

This work was supported by the InnoHK program.

References

Agrawal, A.; Raskar, R.; Nayar, S. K.; and Li, Y. 2005. Removing photography artifacts using gradient projection and flash-exposure sampling. In *SIGGRAPH*.

Chang, Y.; Jung, C.; Sun, J.; and Wang, F. 2020. Siamese Dense Network for Reflection Removal with Flash and No-Flash Image Pairs. *Int. J. Comput. Vis.*, 128(6): 1673–1698.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *European conference on computer vision*, 17–33. Springer.

Chen, W.-T.; Vong, Y.-J.; Kuo, S.-Y.; Ma, S.; and Wang, J. 2024a. RobustSAM: Segment Anything Robustly on Degraded Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4081–4091.

Chen, X.; Jiang, X.; Tao, Y.; Lei, Z.; Li, Q.; Lei, C.; and Zhang, Z. 2024b. Towards Flexible Interactive Reflection Removal with Human Guidance. *arXiv preprint arXiv:2406.01555*.

Chugunov, I.; Shustin, D.; Yan, R.; Lei, C.; and Heide, F. 2023. Neural Spline Fields for Burst Image Fusion and Layer Separation. *arXiv preprint arXiv:2312.14235*.

Dong, Z.; Xu, K.; Yang, Y.; Bao, H.; Xu, W.; and Lau, R. H. 2021a. Location-aware Single Image Reflection Removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4997–5006. Los Alamitos, CA, USA: IEEE Computer Society.

Dong, Z.; Xu, K.; Yang, Y.; Bao, H.; Xu, W.; and Lau, R. W. 2021b. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5017–5026.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Fan, Q.; Yang, J.; Hua, G.; Chen, B.; and Wipf, D. 2017. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*.

Farid, H.; and Adelson, E. H. 1999. Separating reflections and lighting using independent components analysis. In *CVPR*.

Feng, X.; Pei, W.; Jia, Z.; Chen, F.; Zhang, D.; and Lu, G. 2021. Deep-masking generative network: A unified framework for background restoration from superimposed images. *IEEE Transactions on Image Processing*, 30: 4867–4882.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NeurIPS*.

Guo, X.; Cao, X.; and Ma, Y. 2014. Robust separation of reflection from multiple images. In *CVPR*.

Han, B.-J.; and Sim, J.-Y. 2017. Reflection removal using low-rank matrix completion. In *CVPR*.

Hu, Q.; and Guo, X. 2021. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. *Advances in Neural Information Processing Systems*, 34: 24683–24694.

Hu, Q.; and Guo, X. 2023. Single Image Reflection Separation via Component Synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13138–13147.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Kim, S.; Huo, Y.; and Yoon, S.-E. 2020. Single Image Reflection Removal With Physically-Based Training Images. In *CVPR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kong, N.; Tai, Y.; and Shin, J. S. 2014. A Physically-Based Approach to Reflection Separation: From Physical Modeling to Constrained Optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2): 209–221.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Lei, C.; and Chen, Q. 2021. Robust Reflection Removal with Reflection-free Flash-only Cues. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lei, C.; Huang, X.; Qi, C.; Zhao, Y.; Sun, W.; Yan, Q.; and Chen, Q. 2021. A Categorized Reflection Removal Dataset with Diverse Real-world Scenes. *arXiv preprint arXiv:2108.03380*.

Lei, C.; Huang, X.; Zhang, M.; Yan, Q.; Sun, W.; and Chen, Q. 2020. Polarized Reflection Removal With Perfect Alignment in the Wild. In *CVPR*.

Lei, C.; Jiang, X.; and Chen, Q. 2023. Robust reflection removal with flash-only cues in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Levin, A.; and Weiss, Y. 2007. User assisted separation of reflections from a single image using a sparsity prior. *TPAMI*, 29(9): 1647–1654.

Li, C.; Yang, Y.; He, K.; Lin, S.; and Hopcroft, J. E. 2020. Single Image Reflection Removal through Cascaded Refinement. In *CVPR*.

Li, Y.; and Brown, M. S. 2013. Exploiting reflection change for automatic reflection removal. In *ICCV*.

Li, Y.; Liu, M.; Yi, Y.; Li, Q.; Ren, D.; and Zuo, W. 2023. Two-stage single image reflection removal with reflection-aware guidance. *Applied Intelligence*, 1–16.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, Y.-L.; Lai, W.-S.; Yang, M.-H.; Chuang, Y.-Y.; and Huang, J.-B. 2020. Learning to See Through Obstructions. In *CVPR*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lyu, Y.; Cui, Z.; Li, S.; Pollefeys, M.; and Shi, B. 2019. Reflection separation using a pair of unpolarized and polarized images. In *NeurIPS*.

Ma, D.; Wan, R.; Shi, B.; Kot, A. C.; and Duan, L.-Y. 2019. Learning to Jointly Generate and Separate Reflections. In *ICCV*.

Niklaus, S.; Zhang, X.; Barron, J. T.; Wadhwa, N.; Garg, R.; Liu, F.; and Xue, T. 2021. Learned Dual-View Reflection Removal. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3712–3721. Los Alamitos, CA, USA: IEEE Computer Society.

Patrick, W.; Orazio, G.; Jinwei, G.; and Jan, K. 2018. Separating Reflection and Transmission Images in the Wild. In *ECCV*.

Rui, L.; Simeng, Q.; Guangming, Z.; and Wolfgang, H. 2020. Reflection Separation via Multi-bounce Polarization State Tracing. In *ECCV*.

Shih, Y.; Krishnan, D.; Durand, F.; and Freeman, W. T. 2015. Reflection removal using ghosting cues. In *CVPR*.

Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; and Jorge Cardoso, M. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, 240–248. Springer.

Sun, C.; Liu, S.; Yang, T.; Zeng, B.; Wang, Z.; and Liu, G. 2016. Automatic Reflection Removal using Gradient Intensity and Motion Cues. In *ACM MM*.

Wan, R.; Shi, B.; Duan, L.-Y.; Tan, A.-H.; and Kot, A. C. 2017. Benchmarking single-image reflection removal algorithms. In *ICCV*.

Wan, R.; Shi, B.; Li, H.; Hong, Y.; Duan, L.-Y.; and Kot, A. C. 2022. Benchmarking single-image reflection removal algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1424–1441.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

Wei, K.; Yang, J.; Fu, Y.; Wipf, D.; and Huang, H. 2019a. Single Image Reflection Removal Exploiting Misaligned Training Data and Network Enhancements. In *CVPR*.

Wei, K.; Yang, J.; Fu, Y.; Wipf, D.; and Huang, H. 2019b. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8178–8187.

Wen, Q.; Tan, Y.; Qin, J.; Liu, W.; Han, G.; and He, S. 2019. Single Image Reflection Removal Beyond Linearity. In *CVPR*.

Xue, T.; Rubinstein, M.; Liu, C.; and Freeman, W. T. 2015. A computational approach for obstruction-free photography. *ACM Trans. Graph.*, 34(4): 79:1–79:11.

Yang, J.; Gong, D.; Liu, L.; and Shi, Q. 2018. Seeing Deeply and Bidirectionally: A Deep Learning Approach for Single Image Reflection Removal. In *ECCV*.

Yang, Y.; Ma, W.; Zheng, Y.; Cai, J.-F.; and Xu, W. 2019. Fast Single Image Reflection Suppression via Convex Optimization. In *CVPR*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, H.; Xu, X.; He, H.; He, S.; Han, G.; Qin, J.; and Wu, D. O. 2020. Fast User-Guided Single Image Reflection Removal via Edge-Aware Cascaded Networks. *IEEE Transactions on Multimedia*, 22: 2012–2023.

Zhang, X.; ; Ng, R.; and Chen, Q. 2018. Single image reflection separation with perceptual losses. In *CVPR*.

Zheng, Q.; Shi, B.; Chen, J.; Jiang, X.; Duan, L.-Y.; and Kot, A. C. 2021. Single image reflection removal with absorption effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13395–13404.

Zhong, H.; Hong, Y.; Weng, S.; Liang, J.; and Shi, B. 2024. Language-guided Image Reflection Separation. arXiv:2402.11874.

Zhu, Y.; Fu, X.; Jiang, P.-T.; Zhang, H.; Sun, Q.; Chen, J.; Zha, Z.-J.; and Li, B. 2024. Revisiting Single Image Reflection Removal In the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25468–25478.