DL2G: Degradation-guided Local-to-Global Restoration for Eyeglass Reflection Removal

Zhilv Yi[1*], Xiao Lu[2*], Hong Ding[3], Jingbo Hu[1], Zhi Jiang[1], Chunxia Xiao[1†]

[1]School of Computer Science, Wuhan University, Wuhan, Hubei, China
[2]College of Engeering and Design, Hunan Normal University, Changsha, Hunan, China
[3]Guangxi University of Finance and Economics, Nanning, Guangxi, China.

{yizhilv, 2023282110107, zzz1203685136, cxxiao}@whu.edu.cn,
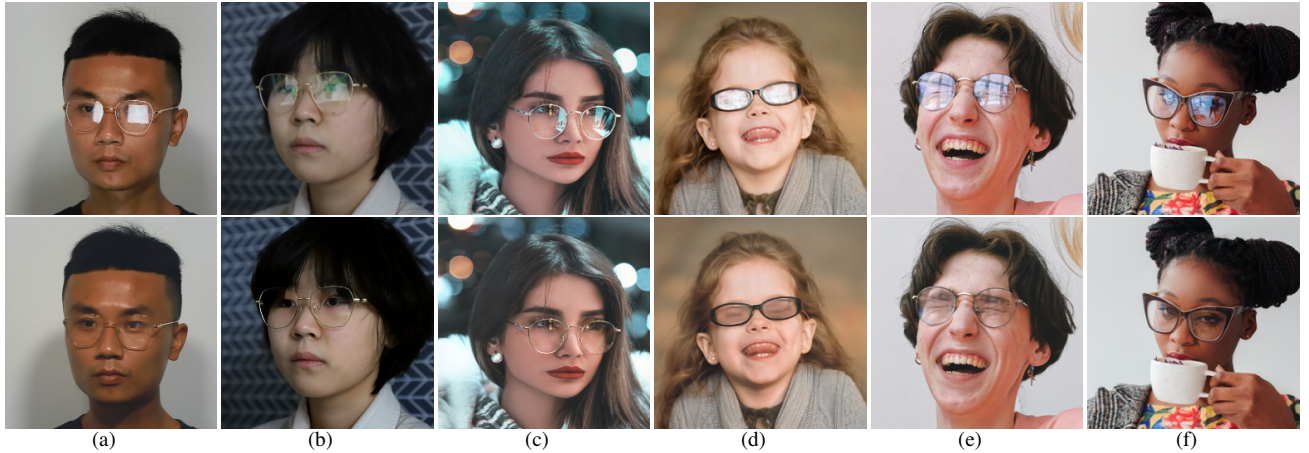luxiao@hunnu.edu.cn, dhong20123@126.com

Figure 1. Examples of our DL2G for eyeglass reflection removal. The top row shows some cases with different reflection intensities, colors, and shapes. The bottom row presents our results. (a), (b) are captured in the lab with controlled lighting conditions. (c) and (d) are captured in the outdoor environment with uncontrolled lighting conditions. (c), (d), (e) and (f) are collected from the Internet.

Abstract

Eyeglass reflection removal can restore the texture information in the reflection destructed eye area, which is meaningful for various tasks on the facial images. It is still challenging to correctly eliminate reflections, reasonably restore the lost contents, and guarantee that the final result has a consistent color and illumination with the input image. In this paper, we introduce a Degradation-guided Local-to-Global (DL2G) restoration framework to address this problem. We first propose a multiplicative reflection degradation model, which is used to alleviate reflection degradation to obtain a preliminary result. Then, in the local details restoration stage, we propose a local structure-aware diffusion model to learn the true distribution of texture details in the eye area. This helps in recovering lost contents in the regions of heavy degradation where the background is invisible. Finally, in the global consistency refinement stage, we utilize the input image as a reference image to generate the final result that is consistent with the input image in color and illumination. Extensive experiments demonstrate that our method can improve the effect of reflection removal and generate results with more reasonable semantics, exquisite details, and harmonious illumination.

1. Introduction

Reflection is a common optical phenomenon observed in natural scenes. For portrait images with eyeglasses, reflections on the eyeglasses often obscure the texture detail of the eyes, even resulting in the lost of content information. Therefore, removing reflections can improve the quality and

---

visibility of the eyes in the image [26]. This is of great importance for visual tasks related to facial image such as facial recognition [22] and facial landmark detection [4].

Different light sources and lens materials will cause the reflection in the real world to show different intensities, shapes, and colors, as shown in Fig. 1. Under indoor conditions, eyeglass reflections are mainly caused by artificial light sources and usually appear as simple geometric shapes with a single color. Under outdoor conditions, eyeglass reflections are mainly caused by natural light, with more complex shapes and more diverse colors. In addition, weak reflection often results in a change in color and brightness information. However, strong reflection can cause the loss of texture information (as shown in Fig. 1 (b)). Thus, it is still a great challenge to correctly remove reflections, reasonably restore the lost contents, and ensure that the generated result has consistent color and illumination with the input image.

In recent years, with the emergence of large-scale reflection datasets [7, 8, 16, 37, 39], many deep learning-based methods have been proposed to remove reflections in both natural scene [11–13, 16, 24, 30, 31, 33, 37] images and eyeglass images [12]. However, these methods struggle to restore local details of the eyes without introducing artifacts and generalize to reflections with different intensities, due to the lack of specific degradation prior to guide the learning process and attention to local structures of the eyes.

Existing degradation prior-guided reflection removal studies focus on separating the reflection layer from the images taken through glass. They consider the degraded image as a blend of the background scene behind the glass and the reflection scene in front of the glass. Consequently, existing methods model the degraded image as a linear [15, 27, 33] or non-linear combination [5] of a reflection image and a transmission image (background image). However, they can not model the eyeglass reflection degradation problem appropriately. Since eyeglass reflections usually only degrade a very small portion of the image, the degraded region can only present local information of the reflection scene in front of the eyeglass. It is challenging and unnecessary to use this local information to estimate the reflection layer. More importantly, separating the reflective layer from the degraded image is not helpful for restoring the lost content information in the strong reflection areas.

Addressing these problems, we construct a Degradation-guided Local-to-Global restoration framework (DL2G) for eyeglass reflection removal (**ERR**). First, we consider that the eyeglass reflection image is formed by applying an illumination and color change surface (degradation map) to the background image, and propose a multiplicative degradation model. Accordingly, we train a simple degradation estimation module (DEM) to estimate the degradation map to alleviate the degradation preliminarily. Then, we propose a two-stage network to recover lost information and

enhance illumination consistency in a local-to-stage manner. Considering that the degradation model is not effective for content restoration in strong reflective regions, we introduce the local sampling scheme to a conditional diffusion model to enhance the model's ability to discern and interpret the local structures of eyes. This is helpful for recovering local contents without introducing artifacts. Finally, we present the global consistency refinement module to incorporate the features of non-eyeglass areas in the input image into the final result, so as to ensure the illumination consistency with the input image. Extensive experimental results on the images captured in the illumination-controlled environment and in the wild all demonstrate that our method can generate reflection-free results with reasonable semantics, exquisite details, and harmonious illumination.

In summary, our main contributions are as follows.

(1) We propose a multiplicative degradation prior model specific to the ERR problem to alleviate the reflection degradation to get a preliminary result.

(2) To restore lost information in areas with high degradation, we train a local structure-aware diffusion model (LSDM) by introducing the local structure sampling strategy for discerning and interpreting local structures. We also propose to incorporate the features of non-eyeglass areas in the input image into the final result to ensure global illumination consistency.

(3) We provide 2,000 pairs of images with more complex reflections as a complement to ReyeR [39], and construct the ReyeR+ dataset. Results on ReyeR+ and other images captured in the wild demonstrate the effectiveness and generalization of our method.

2. Related Work

Reflection Degradation Prior Model. Several natural scene reflection removal studies have proposed reflection degradation prior to separate the reflection scene (reflection image) from the background scene (transmission image). Levin *et al.* [15] proposed that the degraded image can be decomposed as a reflection image and a transmission image. Some approaches [27, 33] further introduced scalars to combine the two components to obtain a more flexible form. Since simple combinations often fail in situations like overexposure [31], [5] introduced an alpha-matting matrix in reflection models. Hu *et al.* [11] proposed a more generic reflection model by introducing a residual term as a non-linear combination of the reflection image and the transmission image. However, it is difficult to separate the reflection scene from the small area of the degradation image on the eyeglass. Besides, the background and reflection are highly mixed, the reflection component dominates the mixture image, and the background is almost invisible, which increases the restoration of background image further.

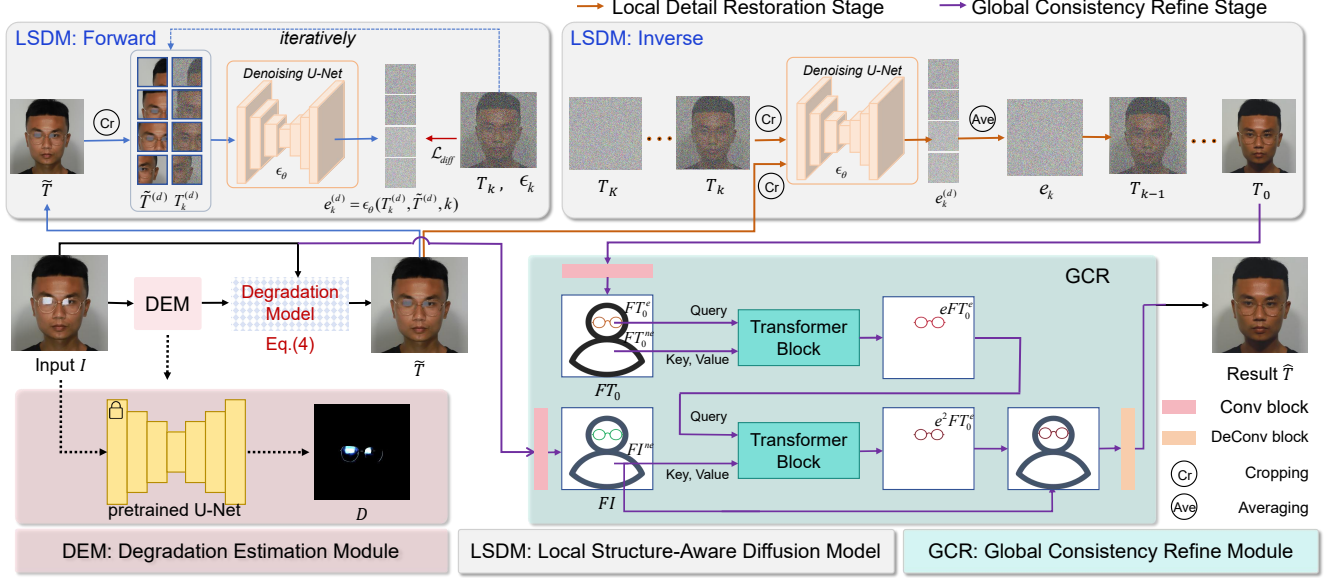**Learning-based Reflection Removal Methods.** Recent

Figure 2. Framework fo our DL2G. Given the input image I, the degradation estimation module (DEM) estimates the degradation map D, obtaining the preliminary result $\widetilde{T}$ according to the proposed degradation model. Then, we use $\widetilde{T}$ as condition to train a local structure-aware diffusion model (LSDM) to restore the lost contents and outputs the result T_0. Finally, the global consistency refinement (GCR) module incorporates the non-eye features of I into the final result $\hat{T}$ to achieve illumination consistency with the input image.

methods mainly adopt deep learning to solve this problem [1–3, 5, 6, 11, 13, 18, 20, 24, 29]. Li *et al.* [16] employed LSTM for iterative refinement of the reflection removal results. Kim *et al.* [13] proposed to transform the reflection images back to their initial states to separate the reflection layer and the transmission layer. Hu *et al.* [11] employed residual term learning to guide reflection separation. Watanabe *et al.* [29] utilized autoencoder and U-Net architecture for ERR. These methods focus on separating the reflection from the degradation image, neglecting the restoration of lost contents in strong reflection regions. Zou *et al.* [39] proposed learning to eliminate weak reflection and restore lost contents jointly. However, the model is not robust to complex reflections, since it cannot learn the true distribution of local details of eyes.

Diffusion Model for Image Restoration. In recent years, diffusion models have achieved excellent results in many vision tasks, such as inpainting [19], portrait highlight removal [38], shadow removal [9], foreground relighting [35], and human reconstruction [17]. Guo *et al.* [9] proposed to embed the degradation prior into the diffusion sampling procedure for shadow removal. Yi *et al.* [34] proposed to formulate the low-light enhancement task as a paradigm of decomposition and image generation based on the Retinex theory. However, these methods mainly use the diffusion model to restore the images globally. Unlike them, we train a details-aware diffusion model to focus on the local contents restoration.

3. Method

Fig.2 illustrates the overview of our DL2G framework. Given an input image I and its corresponding ground-truth image T, we first pre-train a simple U-Net to estimate the degradation map D, which is then used to obtain the initial result $\widetilde{T}$ with the proposed degradation prior model. Then, in the local detail restoration stage, we use $\widetilde{T}$ as condition to train a local structure-aware diffusion model (LSDM) to generate the contents restored result T_0 with the local sampling technique. This benefits to enhance the model's ability to discern and interpret the local structures of eyes. Finally, the global consistency refinement (GCR) module takes T_0 and I as inputs, and fuses the features of eye in T_0 and the features of non-eye in I to generate the illumination consistent result $\hat{T}$.

3.1. Degradation-guided Reflection Alleviation

Given the reflection-degraded image I, many models have been proposed to model the physical mechanism of reflection formation. Hu *et al.* proposed the general form, and most of existing models can be seen as a special case of it:

$$I = T + R + \Phi(T, R) \tag{1}$$

where T and R are transmission and reflection layer, respectively, and $\Phi(T, R)$ can model the residue in a group of specific situations. However, this additive mixing model cannot model the ERR problem appropriately. As we ana-
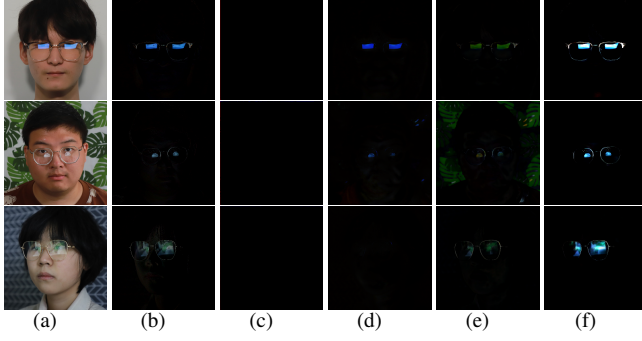
Figure 3. Comparison with different reflection degradation models. (a) I, (b) $R = I - T$, (c) R in IBCLN [16], (d) R in the DSRNet [11], (e) $R + \Phi(T, R)$ in the DSRNet [11], (f) D in our reflection degradation model.

lyzed above, the localized eyeglass reflection makes it hard to separate the reflection scene from the background image.

Multiplicative reflection degradation model. In this paper, inspired by the Retinex theory [14], we consider that the eyeglass reflection image is formed by applying an illumination and color change surface (degradation map) D to the transmission image T as follows:

$$I = T \circ D, \quad (2)$$

where "$\circ$" denotes the element-wise multiplication. With this prior, we can model reflections with change intensity and color. To estimate the degradation map D, we use the simple U-Net to train the Degradation Estimation Module (DEM) with the ground truth $D_T = \frac{I}{T+\eta}$, where η is a decimal used to prevent division by zero ($\eta = 1 \times e^{-4}$ in this paper). The training loss for the degradation map is:

$$\mathcal{L}_D = \| D - D_T \|_F^2, \quad (3)$$

where "$\| \cdot \|_F$" denotes the Frobenious-norm.

We compare different reflection degradation models by visualizing the reflection image R and our degradation map in Fig. 3. It can be seen that the additive degradation priors cannot model the eyeglass reflection well, and our multiplicative model can accurately estimate the degraded map.

According to Eq. (2), we can obtain the initial result of the reflection removal $\widetilde{T}$ by the inverse operation:

$$\widetilde{T} = I./(D + \eta) \quad (4)$$

where "$./$" denotes the element-wise division. We also report the initial result $\widetilde{T}$ in Fig. 4 to show the effect of the degradation model. It can be seen that our model is effective in the weak reflection regions (as seen in Fig. 4(c)). As for the strong reflection regions shown in Fig. 4(f), the degradation model can only eliminate the reflections to a certain extent but cannot restore the lost information. Therefore, this

stage allows us to remove weak degradation, since it cannot restore the heavy degradation regions where the background is almost invisible.

(a) Input (b) T (GT) (c) $\widetilde{T}$ (d) Input (e) T (GT) (f) $\widetilde{T}$
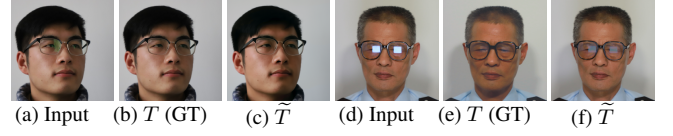
Figure 4. The degradation-guided elimination stage can remove weak reflection ((a) and (c)), but cannot restore the lost content in strong reflection image ((d) and (f)).

3.2. Local Detail Restoration Stage

After the degradation-guided reflection alleviation, we need to restore the lost texture in the strong reflection regions. In this regard, we can learn the latent distribution of eyes with a conditional diffusion-based [23] model. It learns a conditional reverse process $p_\theta(T_{0:K}|\widetilde{T})$:

$$p_\theta(T_{0:K}|\widetilde{T}) = p(T_K) \prod_{k=1}^{K} p_\theta(T_{k-1}|T_k, \widetilde{T}), \quad (5)$$

such that the sampled image has high fidelity to the distribution of T conditioned on the given initial result $\widetilde{T}$. This can be trained by introducing noise to the reflection-free image T_0, transitioning it to a noisy state T_k over k steps through $T_k = \sqrt{\bar{\alpha}_k}T_0 + \sqrt{1 - \bar{\alpha}_k}\epsilon$, where $\alpha_k = 1 - \beta_k$, $\bar{\alpha}_k = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ($\mathbf{I}$ is the identity matrix). The denoiser ϵ_θ is used to predict the noise map e_k as follows:

$$e_k = \epsilon_\theta(\sqrt{\bar{\alpha}_k}T_0 + \sqrt{1 - \bar{\alpha}_k}\epsilon, \widetilde{T}, k). \quad (6)$$

Then, the denoiser can be trained to predict the noise ϵ and restore the image via the following loss:

$$\mathcal{L} = \mathbb{E}_{T_0,k,\epsilon} \| e_k - \epsilon_k \|_F^2. \quad (7)$$

However, we found that directly training such a model did not result in accurate restoration images. Instead, the model produces images with some artifacts in the strong reflection regions. We speculate that this discrepancy arises because eyeglass reflection usually only occupies a very small part of the whole facial image, and using the whole facial image as a condition to guide the training of diffusion model cannot make the model focus on the local structure.

To enable the diffusion model to be aware of local structures, we propose the local structure-aware diffusion model (LSDM) by introducing the local structure sampling strategy. Specifically, we employ a sliding window of size $p \times p$ with stride $s \times s$ to segment the reflection-free image $T_0 \in \mathbb{R}^{H \times W \times 3}$ and the corresponding conditional image $\widetilde{T} \in \mathbb{R}^{H \times W \times 3}$ into D patches, denoted as $T_k^{(d)} \in \mathbb{R}^{p \times p \times 3}$ and $\widetilde{T}^{(d)} \in \mathbb{R}^{p \times p \times 3}$, where $d = 1, 2, \cdots, D$, $p = 64$,

Algorithm 1 Local structure-aware diffusion training.
___
Input: Initial eyeglass reflection result $\widetilde{T}$, reflection-free image T_0, dictionary of D with overlapping patch locations.

1: **while** not converged **do**
2:     $k \sim \text{Uniform}\{1, \cdots, K\}$
3:     $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4:     **for** $d \in [1, D]$ **do**
5:        $T_0^{(d)} = Crop(P_d \circ T_0)$, $\widetilde{T}^{(d)} = Crop(P_d \circ \widetilde{T})$, and $\epsilon^{(d)} = Crop(P_d \circ \epsilon)$ (P_d is the mask of the d_{th} patch in the image.)
6:        $e_k^{(d)} = \epsilon_\theta(\sqrt{\bar{\alpha}_k}T_0^{(d)} + \sqrt{1 - \bar{\alpha}_k}\epsilon^{(d)}, \widetilde{T}^{(d)}, k)$
7:        Perform gradient decent steps on $\nabla_\theta \mathcal{L}_{diff}$
8:     **end for**
9: **end while**
10: **return** ϵ_θ
___

Algorithm 2 Inference procedure of LSDM.
___
Input: Initial eyeglass reflection result $\widetilde{T}$, diffusion model ϵ_θ, dictionary of D overlapping patch locations.

1: $T_K \sim \text{Uniform}\{1, \cdots, K\}$
2: **for** k = K, ..., 1 **do**
3:     $\Omega_k = 0$ and $M = 0$
4:     **for** $d \in [1, D]$ **do**
5:        $T_k^{(d)} = Crop(P_d \circ T_k)$, $\widetilde{T}^{(d)} = Crop(P_d \circ \widetilde{T})$
6:        $\Omega_k = \Omega_k + P_d \circ \epsilon_\theta(T_k^{(d)}, \widetilde{T}^{(d)}, k)$
7:        $M = M + P_d$
8:     **end for**
9:     $\Omega_k = \Omega_k./M$
10:     $T_{k-1} = \frac{1}{\alpha_k}(T_k - \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}}\Omega_k) + \sqrt{1 - \alpha_k}\epsilon_k$ ($\epsilon_k \sim \mathcal{N}(0, \mathbf{I})$)
11: **end for**
12: **return** T_0
___

$s = 16$, and $T_k^{(d)}$ is the reflection-free image with Gaussian noise ϵ_k in timestep k. Then, the learned conditional reverse process in Eq. (5) can be defined as:

$$p_\theta(T_{0:K}^{(d)})|\widetilde{T}^{(d)}) = p(T_K^d) \prod_{k=1}^{K} p_\theta(T_{k-1}^{(d)}|T_k^{(d)}, \widetilde{T}^{(d)}). \quad (8)$$

The denoiser in Eq. (6) and the loss function in Eq. (7) can be rewritten as:

$$e_k^{(d)} = \epsilon_\theta(\sqrt{\bar{\alpha}_k}T_0^{(d)} + \sqrt{1 - \bar{\alpha}_k}\epsilon^{(d)}, \widetilde{T}^{(d)}, k), \quad (9)$$

$$\mathcal{L}_{diff} = \mathbb{E}_{T_0^{(d)}, k, \epsilon^{(d)}} \parallel e_k^{(d)} - \epsilon_k^{(d)} \parallel_F^2 . \quad (10)$$

The training procedure for our local structure-aware diffusion model is presented in Algorithm 1.

In the reverse procedure, as presented in Algorithm 2, we process the overlapped regions at every timestep k to ensure a uniform effect across the overlapped areas when reconstructing the whole image. This prevents the formation of merging artifacts that typically occur when patches are processed independently. With the local structure-aware diffusion model, we can obtain a seamless and artifact-free restoration image.

3.3. Global Consistency Refinement Stage

The result T_0 obtained by the LSDM is more refined in detail but sometimes exists illumination deviations from the input image, therefore, we further design a global consistency refine (GCR) module to refine the illumination with the help of reference image I. Then we can get the final result $\hat{T}$ harmonious with the input image. The GCR module consists of two convolutional blocks for image feature extraction, two transformer blocks [25] for feature fusion, and a deconvolutional block for image generation.

Specifically, the GCR module takes T_0 and the image I as input, and performs feature extraction to get the feature maps FT_0 and FI through a convolutional block. Then, we can obtain the feature in and out-side the eye area of these two images, *i.e.*, FT_0^e, (FT_0^{ne}, and FI^{ne}), respectively, through the following operations:

$$FT_0^e = Flatten(FT_0 \circ M),$$
$$FT_0^{ne} = Flatten(FT_0 \circ (1 - M))), \quad (11)$$
$$FI^{ne} = Flatten(I \circ (1 - M))),$$

where M represents the eye area mask, which can be obtain by a glass detector[1]. We first use FT_0^e as query and FT_0^{ne} as key and value to enhance the feature of eyes with the global features of T_0. Then, we can get the enhanced eye feature eFT_0^e. After that, we further use eFT_0^e as query and the non-eye area features FI^{ne} as key and value, and feed them to the transformer block for feature fusion, so that the illumination and tone are as consistent as possible with the input image I. Finally, we combine the enhanced result $e^2FT_0^e$ with the non-eye area content FI_{ne} of the input image to obtain the final optimized result $\hat{T}$.

In the global consistency refine stage, we use the least square loss for training the GCR module:

$$\mathcal{L}_g = \left\|\hat{T} - T\right\|_F^2 . \quad (12)$$

4. Experiments

4.1. Implementation Details

Our network is implemented in PyTorch on a NVIDIA GeForce 3090 card. We train the proposed modules separately. The training epoches of all three modules are
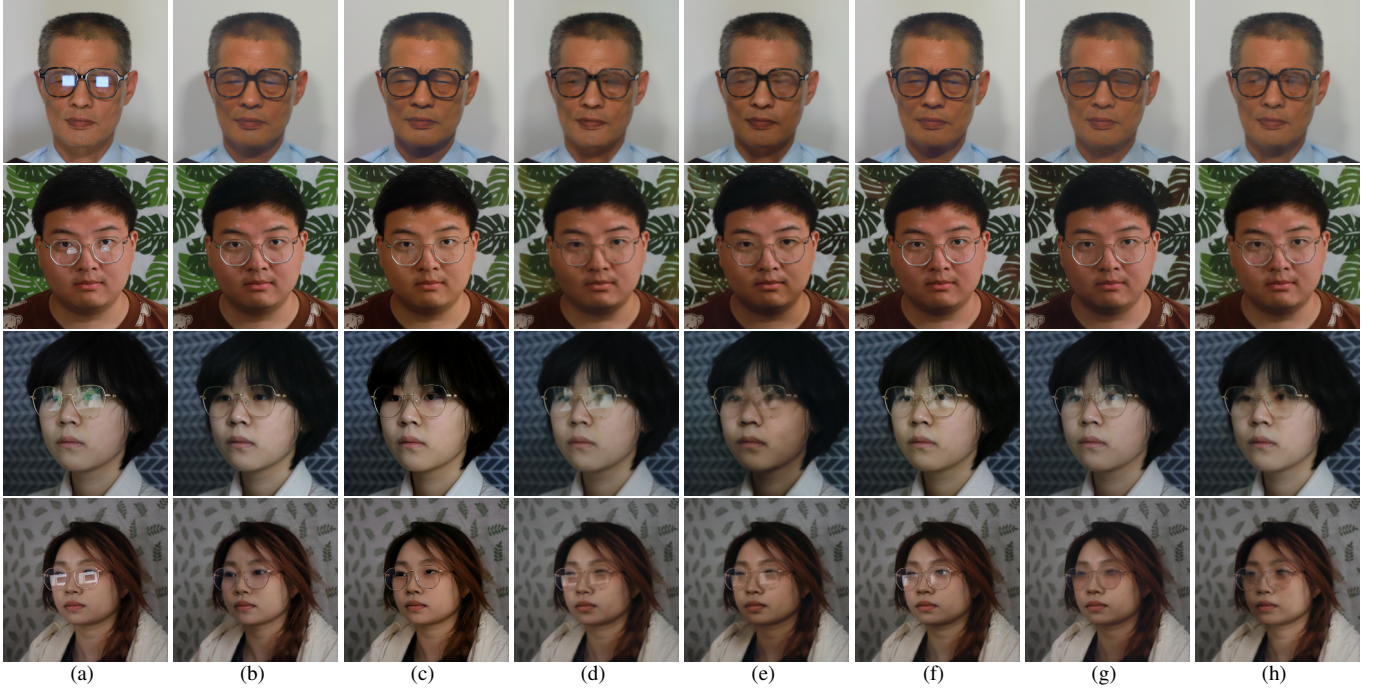
___
[1]https://github.com/mantasu/glasses-detector

Figure 5. Qualitative comparison results on ReyeR+, the first two rows are from ReyeR, and the last two rows are our supplemented images. (a) Input image, (b) GT, (c) Ours, (d) Watanabe *et al.* [29], (e) IBCLN [16], (f) Robust SIRR [24], (g) DSRNet [11], (h) ER^{2}Net [39].

set as 2000. We use Adam optimizer with momentum as (0.9,0.999) to optimize the DEM, LSDM, and GCR modules, and set the initial learning rate to 3×10^{-5}. For the DEM, the recent transformer image-to-image backbone [28]. For the LSDM, we use the similar U-Net architecture as the denoiser ϵ_θ of [21], and initialize the weights of our model with the Kaiming initialization technique [10]. The noise schedule β_k increases linearly from 0.0001 to 0.02. The diffusion step K is set as 1000 for training and 100 for inference. For the GCR module, we use the transformer blocks in [25] for feature fusion. All input images are resized to 256×256.

4.2. Comparison Results on ReyeR+

Dataset. The ReyeR dataset contains 12,328 pairs of high-quality eyeglass reflection images with 11,046 training pairs and 1,282 testing pairs. The images were captured with various materials of eyeglasses, different lighting sources, and different reflection intensities. However, the degradation levels of these images are relatively simple, *e.g.,* small areas and regularly shaped. In this work, we use the same collecting scheme as that in [39] to supplement ReyeR with 2,000 pairs of images, and the new dataset is denoted as ReyeR+ (details can be seen in Table 1). We mainly collect some reflection images with complex foreground scenes and reflection source, in most of which the eyes are completely invisible. The supplemented images are more closely resemble the real-world scenarios of complex reflections under nat-

ural indoor/outdoor lighting conditions, thereby presenting greater challenges.

Table 1. Details comparison of ReyeR and ReyeR+.

Dataset	Participants	Light sources	Eyes invisible(%)	Complex foreground(%)	training set
ReyeR	356	5	12.13	2.01	11,046
ReyeR+	400	20	33.24	10.14	11,846

Metrics. For the labeled ReyeR+, we calculate PSNR, SSIM and LPIPS [36] on the RGB space for evaluation, and for unlabeled images in the wild, qualitative comparisons are provided for visual observation.

Table 2. Quantitative comparison results on the ReyeR+ dataset. The best results are in **bold**, and the second-best results are in underlined.

Methods	PSNR↑	SSIM↑	LPIPS↓
Watanabe *et al.* [29]	31.514	0.857	0.130
ER^{2}Net [39]	<u>34.301</u>	<u>0.935</u>	0.050
IBCLN [16]	32.471	0.882	0.085
Robust SIRR [24]	32.761	0.917	0.073
DSRNet [11]	33.555	0.933	<u>0.046</u>
TSHRNet [8]	20.257	0.879	0.136
Ours	**34.801**	**0.947**	**0.040**

We compare our method with two eyeglass reflection removal method Watanabe *et al.* [29] and ER^{2}Net [39], three advanced image reflection removal methods IBCLN [16],
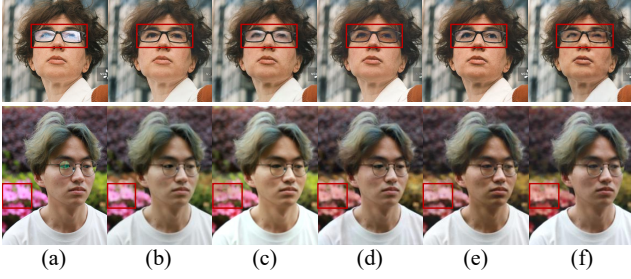
Figure 6. Results on images in the wild. (a) Input image, (b) Ours, (C) IBCLN, (d) DSRNet, (e) Robust SIRR, (f) ER^{2}Net.
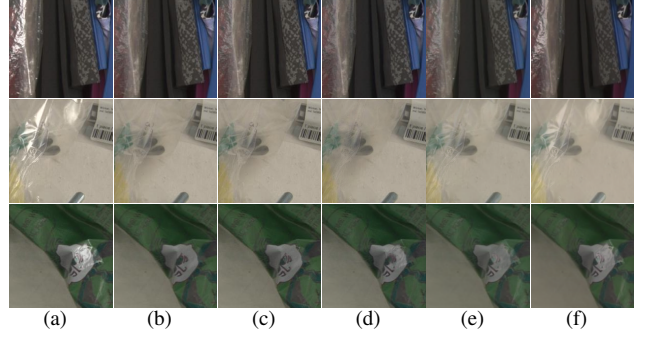

Figure 7. Comparison with specular highlight removal methods. (a) Input; (b) GT; (c) Ours; (d) TSHRNet [8]; (e) JSHDR [7]; (f) SpecularityNet [32].

Table 3. Quantitative comparison results on SHIQ dataset. The best results are marked in **bold**, and the second-best results are in underlined.

Metrics	PSNR↑	SSIM↑
Fu [7]	**34.131**	0.860
Wu [32]	23.420	0.920
Fu [8]	25.575	<u>0.933</u>
Ours	<u>28.145</u>	**0.946**

DSRNet [11] and Robust SIRR [24], and an advanced specular highlight removal method TSHRNet [8]. We train these networks on ReyeR+ dataset with the optimal settings as mentioned in the original published paper. The quantitative comparison results are reported in Table 2. As we can see, our method performs the best on all the three metrics. Fig. 5 shows the qualitative comparison results on images with different reflection intensities, shapes, and colors. For the Watanabe *et al.* [29], IBCLN [16], Robust SIRR [24], DSRNet [11] and ER^{2}Net [39], in the case of reflections with single shape and color, as shown in the first 4 columns of Fig. 5, these methods can remove reflections. However, in the strong reflection regions, the contents recovered by these methods are not detailed enough, and there are distortions and background fading. For reflections with complex patterns, as shown in the last two columns of Fig.5, these methods are difficult to completely remove reflections and restore the contents.

More Results on Images in the Wild. To further verify the robustness and generalization ability of our method, we test our method on some images in the wild. The visualization results are presented in Fig, 6. It can be seen that our method can recover the background image with reasonable details (see the red box in the first row) and consistent illumination (see the red box in the second row), which demonstrates the effectiveness of our method on the images captured in the environment with uncontrolled lighting conditions. More results on the images in the wild can be seen in Fig. 1, and more comparisons on these images are presented in the **Supplementary Material**.

4.3. Results on Specular Highlight Images

Since the eyeglass reflections is similar to specular highlights, we generalize our method to specular highlight removal task. We compare our method with three SOTA specular highlight removal methods, TSHRNet [8], JSHDR [7], and SpecularityNet [32]. Table 3 and Fig. 7 show the qualitative and quantitative comparison results on specular highlight dataset SHIQ [7], respectively. It can be seen that our method can achieve equivalent even higher scores than other current specular highlight removal methods. Even in some results, our method is better at detail processing and restores the lost texture on the object more delicately.

4.4. Ablation Study

We design four variants as follows to verify the effectiveness of each module in our network: (1) **only DEM:** Only using the degradation estimation model for reflection alleviation to see the effectiveness of the degradation model. (2) **only LSDM:** Only using the local structure-aware diffusion model for reflection removal with $\widetilde{T}$ replaced by I to understand the performance of the LSDM module. (3) **DEM + LSDM:** Using the degradation-guided reflection alleviation and the LSDM to get the final result. (4) **DEM + GCR:** Using the degradation-guided reflection alleviation and the global consistency refinement module, with T_0 replaced by $\widetilde{T}$, to get the final result.

All the four variants are trained on the ReyeR+ dataset, and the results are summarized in Table 4. We can observe that all modules can improve the eyeglass reflection removal performance, which indicates the effectiveness of our design. Fig. 8 provides a visual comparison of the ablation variants. Fig. 8(c) shows degradation module is effective to alleviate the degradation, but can not recover the lost contents. LSDM is effective to restore the local details, but there are still some reflection residuals (see Fig. 8(d)). The variant DEM + GCR can not obtain the results with rea-
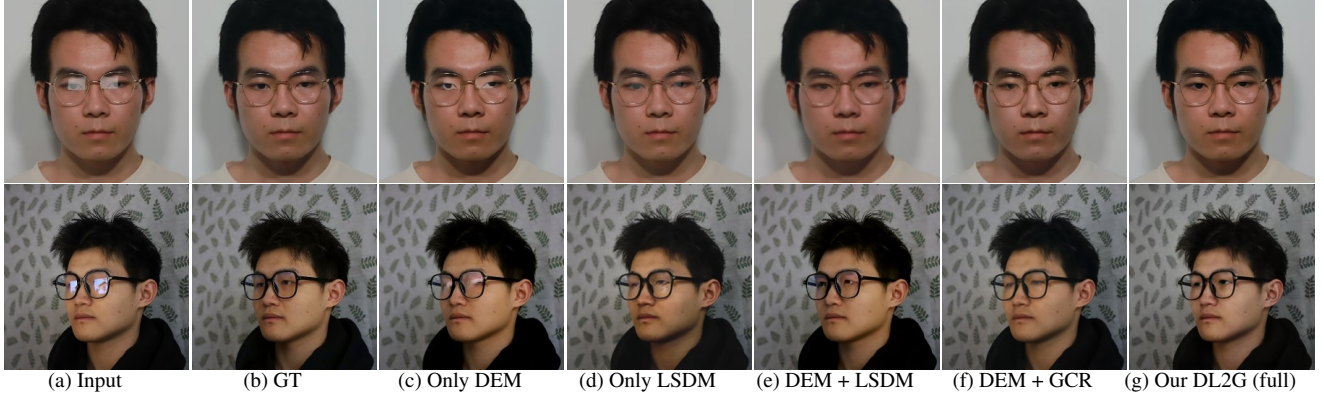
(a) Input   (b) GT   (c) Only DEM   (d) Only LSDM   (e) DEM + LSDM   (f) DEM + GCR   (g) Our DL2G (full)

Figure 8. Visualization of ablation study results.

(a)   (b)   (c)   (d)   (e)   (f)   (g)

Figure 9. Failure cases. (a) Input, (b) Ours, (c) Watanabe *et al.* [29], (d) IBCLN [16], (e) Robust SIRR [24], (f) DSRNet [11], (g) ER^{2}Net [39].

Table 4. Quantitative comparison results in ablation study. The best results are marked in **bold**.

Variants	PSNR↑	SSIM↑	LPIPS↓
only DEM	31.633	0.894	0.128
only LSDM	33.441	0.932	0.051
DEM + LSDM	34.754	0.934	0.044
DEM + GCR	32.490	0.919	0.061
DL2G (full)	**34.801**	**0.947**	**0.040**

sonable details (see Fig. 8(f)), and this can be improved by replacing GCR with LSDM, but it cannot obtain the illumination harmonious result (see Fig. 8(d)). Our DL2G combines the above modules to output high quality results.

Limitations. Our method may fail on the eyeglasses not on eyes (the top row of Fig. 9). The main reason is that the training dataset mainly focus on restoring eyes. However, comparing with other method, our method still can eliminate the reflections to some extent with the help of the guidance of the degradation model. Another case is for eye-

glasses with special material (the bottom row of Fig. 9), the main reason is that our degradation model cannot separate the front scene from the background.

5. Conclusion

In this paper, we propose a DL2G framework for ERR. The proposed multiplicative degradation model is effective to alleviate the reflection, the LSDM module can learn the true distribution of local structures of eyes for reasonable restoration of the texture details, and the global illumination can be adjusted by incorporating the background features of the input image into the final result. Extensive experimental results show that our DL2G framework can significantly improve the performance of reflection removal and reasonably recover the contents in the strong reflection regions.

6. Acknowledgments

References

[1] Ali Amanlou, Amir Abolfazl Suratgar, Jafar Tavoosi, Ardashir Mohammadzadeh, and Amir Mosavi. Single-image reflection removal using deep learning: a systematic review. *IEEE Access*, 10:29937–29953, 2022. 3

[2] Ya-Chu Chang, Chia-Ni Lu, Chia-Chi Cheng, and Wei-Chen Chiu. Single image reflection removal with edge guidance, reflection classifier, and recurrent decomposition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2033–2042, 2021.

[3] Wei-Ting Chen, Kuan-Yu Chen, I-Hsiang Chen, Hao-Yu Fang, Jian-Jiun Ding, and Sy-Yen Kuo. Missing recovery: Single image reflection removal based on auxiliary prior learning. *IEEE Transactions on Image Processing*, 32:643–656, 2022. 3

[4] Savina Colaco and Dong Seog Han. Facial keypoint detection with convolutional neural networks. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 671–674. IEEE, 2020. 2

[5] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5017–5026, 2021. 2, 3

[6] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3238–3247, 2017. 3

[7] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. A multi-task network for joint specular highlight detection and removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7752–7761, 2021. 2, 7

[8] Gang Fu, Qing Zhang, Lei Zhu, Chunxia Xiao, and Ping Li. Towards high-quality specular highlight removal by leveraging large-scale synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12857–12865, 2023. 2, 6, 7

[9] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[11] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13138–13147, 2023. 2, 3, 4, 6, 7, 8

[12] Meiguang Jin, Sabine Süsstrunk, and Paolo Favaro. Learning to see through reflections. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2018. 2

[13] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5164–5173, 2020. 2, 3

[14] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 4

[15] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007. 2

[16] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3565–3574, 2020. 2, 3, 4, 6, 7, 8

[17] Yuanzhen Li, Fei Luo, and Chunxia Xiao. Diffusionfof: Single-view clothed human reconstruction via diffusionbased fourier occupancy field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9525–9534, 2024. 3

[18] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Semantic guided single image reflection removal. *ACM Transactions on Multimedia Computing, Communications and Applications*, 18(3s):1–23, 2022. 3

[19] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 3

[20] BH Prasad, Lokesh R Boregowda, Kaushik Mitra, Sanjoy Chowdhury, et al. V-desirr: Very fast deep embedded single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2390–2399, 2021. 3

[21] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 6

[22] Reena Sharma, Vijay Kumar Sharma, and Arjun Singh. A review paper on facial recognition techniques. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 617–621, 2021. 2

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[24] Zhenbo Song, Zhenyuan Zhang, Kaihao Zhang, Wenhan Luo, Zhaoxin Fan, Wenqi Ren, and Jianfeng Lu. Robust single image reflection removal against adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24688–24698, 2023. 2, 3, 6, 7, 8

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 6

[26] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017. 2

[27] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2018. 2

[28] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 6

[29] Sota Watanabe and Makoto Hasegawa. Reflection removal on eyeglasses using deep learning. In *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 1–4. IEEE, 2021. 3, 6, 7, 8

[30] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019. 2

[31] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019. 2

[32] Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jianwei Guo, Jun Xiao, Xiaopeng Zhang, and Dong-Ming Yan. Single-image specular highlight removal via real-world dataset construction. *IEEE Transactions on Multimedia*, 24:3782–3793, 2021. 7

[33] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the european conference on computer vision (ECCV)*, pages 654–669, 2018. 2

[34] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 3

[35] Ziqi Yu, Jing Zhou, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. Cfdiffusion: Controllable foreground relighting in image compositing via diffusion model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3647–3656, 2024. 3

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[37] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 2

[38] Hongsheng Zheng, Zhongyun Bao, Gang Fu, Xuze Jiao, and Chunxia Xiao. Phr-diff: Portrait highlight removal via patch-aware diffusion model. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025. 3

[39] Wentao Zou, Xiao Lu, Zhilv Yi, Ling Zhang, Gang Fu, Ping Li, and Chunxia Xiao. Eyeglass reflection removal with joint learning of reflection elimination and content inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10266–10280, 2024. 2, 3, 6, 7, 8