PHR-DIFF: Portrait Highlights Removal via Patch-aware Diffusion Model

Hongsheng Zheng[1], Zhongyun Bao[2], Gang Fu[3], Xuze Jiao[1], Chunxia Xiao[1]*

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]School of Computer and Information, Anhui Polytechnic University, Wuhu, China
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
hszheng@whu.edu.cn, tantouxy@163.com, xyzgfu@gmail.com, 2023282110161@whu.edu.cn, cxxiao@whu.edu.cn

Abstract

Portraits often suffer from specular highlights due to factors like skin oiliness, lighting conditions, and shooting angles, which degrade aesthetics and affect downstream tasks. Thus, portrait highlight removal is imperative. Previous methods struggle to remove highlights and achieve high-fidelity restoration of disturbed regions simultaneously. In this work, we propose a novel patch-based diffusion model for this task, named PHR-DIFF. Specifically, in the training, we present a patchify training strategy that divides the portrait into equal-sized patches and performs diffusion on these patches individually. This patchify can extract more compact facial features and reduce training costs. Besides, to learn the global coherence of the face, we propose a patch-residual approach. It encodes the full-resolution highlight-free portrait into latent features, which are further used as residual terms to constrain the forward training. In the sampling, we remove portrait highlights in a patch-wise manner and propose a Patch-Aware Highlight Removal (PAHR) mechanism. PAHR leverages features from non-highlight regions to effectively guide the patch-wise removal of highlight components. Experimental results on multiple public datasets demonstrate that PHR-DIFF removes highlights more cleanly and avoids artifacts.

Introduction

Portrait plays a crucial role in daily life for identity verification, social interaction, and personal presentation. However, portrait inevitably suffers from unsatisfactory highlight interference, which is mainly due to the reflective properties of facial skin or interference from external lighting. As illustrated in Figure 1, highlights obscure the natural radiance and textures of the face and significantly degrade the visual quality of the portrait. Therefore, portrait highlight removal is imperative. This task aims to remove highlight components caused by strong reflections to enhance the quality and usability of the portrait.

Most early highlight removal methods are based on the physical priori knowledge of the image, such as dark channel separation (Kim et al. 2013), color statistical analysis (Li et al. 2017; Tan and Ikeuchi 2005; Mallick et al. 2006; Son et al. 2020), and brightness statistics (Todd, Norman, and Mingolla 2004). However, in practical applications, obtaining such prior information is challenging, making these methods often unsatisfactory and inefficient. Recently, with the advancement of deep learning, highlight removal methods such as GANs (Fu et al. 2021; Muhammad et al. 2020; Liang et al. 2021; Wu et al. 2021) and Transformer (Wu et al. 2023) have started to emerge. However, due to factors such as the texture characteristics of facial skin and the complexity of geometric contours, existing portrait highlight removal methods often struggle to completely remove specular highlights, typically resulting in residual artifacts and distortions.

Figure 1: Visual examples from the CelebA dataset and our removal results. (a) The specular highlight disrupts the aesthetic of the portrait and obscures the details of the facial texture. (b) The overall aesthetic quality of the portrait is effectively enhanced by our highlight removal method.

In fact, specular highlights of the portrait are often concentrated in specific areas or distributed sporadically. Regardless of their distribution, the underlying features in the original facial (highlight-free) image should exhibit overall coherence in texture and skin tone. This observation greatly motivates us to leverage the features in non-highlight regions to guide the removal of the highlight components. Besides, the diffusion model has recently demonstrated excellent capabilities in various image restoration tasks (Guo et al. 2023a; Ding et al. 2023; Ye et al. 2024; Bar-Tal et al. 2023). Based on this insight, we propose a novel patch-based diffusion model for this task, termed PHR-DIFF. Our PHR-DIFF not only removes specular highlights in the portrait but also enables high-realistic detail restoration of disturbed areas (e.g., textures, skin tones, wrinkles).

*Chunxia Xiao is the corresponding author.

PHR-DIFF is a patch-based conditional diffusion model. Specifically, it takes the highlight-free portrait and highlight portrait as the training inputs. Here the highlight portrait is used for the condition (Guo et al. 2023a). Inspired by (Ding et al. 2023; Wang et al. 2024), the highlight-free portrait is divided into fixed-size patches and then forward noise is added to these patches respectively. Additionally, we perform position embedding for each patch. This patchify approach extracts more compact and subtle local features and improves training efficiency.

However, a notable drawback is the inability to capture the global facial dependencies, which can lead to incoherence in the removal results. To address this issue, we further propose a patch-residual training strategy. It encodes the full-resolution highlight-free portrait into latent features and then uses these features as residual terms to constrain the diffusion training forward to a specific distribution. In the sampling, it starts with pure noise distribution and removes the highlight components by patch-wise denoising. We present a Patch-Aware Highlight Removal (PAHR) module, which utilizes features in non-highlight areas to guide the removal of highlight components, ensuring the cleaning removal and avoiding artifacts.

Overall, our contributions can be summarised as follows:

- We present a tailored patch-based diffusion model for portrait highlight removal, named PHR-DIFF, which utilizes the generative capabilities of the diffusion model to achieve high-fidelity removal results.
- We propose a Patch-Residual Training Strategy (PRTS) for the PHR-DIFF, which significantly reduces the training costs and preserves facial coherence effectively.
- We develop a Patch-Aware Highlight Removal (PAHR) module for sampling, which leverages the information from non-highlight regions to effectively guide the restoration of highlight areas.

Related Work

Specular highlights removal is a fundamental task in computer vision. Traditional works usually rely on physical characteristics and geometric relationships to remove highlights by analyzing the color, brightness, and normal information of the image. Techniques in this category include color-space transformation (Guo et al. 2014), color statistical analysis (Li et al. 2017; Tan and Ikeuchi 2005; Mallick et al. 2006; Son et al. 2020), polarization analysis (Feng et al. 2023; Li et al. 2015b) and reflection separation (Tan and Ikeuchi 2005; Li et al. 2015a). These methods often require a significant amount of hand-crafted prior information. However, in practice, obtaining such information accurately can be challenging.

Recently, learning-based methods have made significant progress in image specular highlight removal. (Wu et al. 2021, 2020) employed generative adversarial networks (GANs) to remove the image highlight and constructed a real natural object highlight dataset utilizing polarization properties. (Yi et al. 2020) leveraged a multi-view image set for unsupervised intrinsic decomposition and highlight separation. (Fu et al. 2021) constructed a dataset for highlight removal and integrated the detection and removal of highlights into a multi-level network. (Fu et al. 2023a) proposed a three-stage network that solves the problem of refinement and tonal deviation in the highlight removal process. (Wu et al. 2023) achieved the specular highlight detection and removal by jointly optimizing a U-Net and Transformer architecture. (Fu et al. 2024) proposed a patch-level bidirectional refinement highlight detection network.

Particularly, some works have been conducted on highlight removal in portrait, (Li et al. 2017, 2015a) performed highlights removal based on the physical and statistical properties of facial skin. (Wang et al. 2021) trained a GANs for face highlight removal on a synthetic dataset and fine-tuned this model on in-the-wild images. To meet the training requirements, (Liang et al. 2021) constructed a high-resolution face dataset for this task and utilized GAN for highlight removal. (Muhammad et al. 2020) also presented a Spec-Face dataset designed to remove high-intensity specular reflections from low chromaticity face images. (Su et al. 2022) designed a lightweight network for removing highlights from multi-view facial images based on Lambertian consistency. Despite efforts from various approaches, these methods often exhibit limitations in handling complex facial contours, blurred textures, and strong highlights.

Diffusion generative models have recently become prevalent in image restoration (Sohl-Dickstein et al. 2015). These models add Gaussian noise to the input data during the forward diffusion process and then learn to reverse this process during sampling to recover the original data. Denoising diffusion probabilistic models (DDPMs) (Ho et al. 2020; Quinn and Dhariwal 2021) utilized a stochastic inverse generative process to progressively denoise images generated from pure noise. Denoising diffusion implicit models (DDIMs) (Song et al. 2020) introduced a deterministic inverse generation process, accelerating sampling and yielding higher-quality generated images. Such models have been widely employed in various visual tasks, including inpainting (Lugmayr et al. 2022; Xie et al. 2023), super-resolution (Rombach et al. 2022; Sauer et al. 2024), shadow removal (Guo et al. 2023a,b), weather restoration (Ozan and Robert 2023) and face relighting (Ponglertnapakorn et al. 2023). (Yu et al. 2024; Zhou et al. 2024) used diffusion models to achieve image harmonization and generate more realistic results.

Methodology

Figure 2(a) illustrates the overview of our proposed PHR-DIFF. PHR-DIFF is a patch-based conditional diffusion model that utilizes the patchify strategy during both training and sampling. In the training, the highlight-free portrait is divided into equal-sized patches for forward diffusion and the highlight portrait is used as the condition. To learn the facial global representation, we encode the full-resolution highlight-free portrait into latent features and then use these features as residual terms to constrain the training, presented in Figure 2(b). In the sampling, we start with pure noise and remove the highlight components by patch-wise denoising. The Patch-Aware Highlight Removal (PAHR) module discriminates between highlight and non-highlight areas by the highlight mask. The non-highlight features are then used to

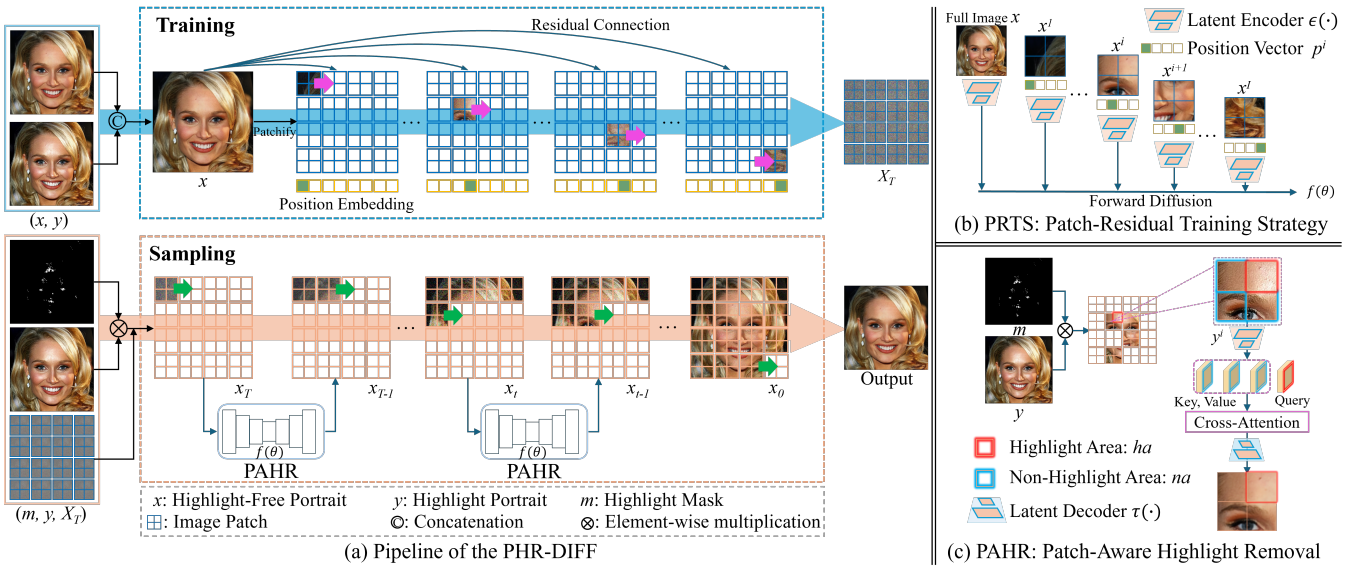


Figure 2: (a) Overview of the proposed PHR-DIFF. In the training phase, PHR-DIFF takes the highlight-free portrait x and highlight portrait y as inputs. x is divided into fixed-size patches for forward diffusion, and y is used as the condition. During the sampling phase, highlight portrait y, highlight mask m, and pure noise distribution X_T are used as inputs for patch-wise highlight removal. (b) To learn a global representation of the portrait, we introduce a Patch-Residual Training Strategy (PRTS). It encodes the full-resolution x into latent features and then uses these features as residual terms to constrain the diffusion toward a specific distribution. (c) To fully leverage the features from non-highlight areas to guide the removal of the highlight, we propose a Patch-Aware Highlight Removal (PAHR) module.

guide removing the highlight components, shown in Figure 2(c). The following sections will provide a detailed explanation of the motivation and formulation behind PHR-DIFF.

Patch-Residual Training for PHR-DIFF. As illustrated in Figure 2(a), our PHR-DIFF takes the highlight-free portrait x and highlight portrait y as inputs. Instead of performing diffusion training on the full-resolution image x, we first divide it into patches. This patchify strategy can learn more compact facial feature representations and enhance training efficiency (Wang et al. 2024; Ding et al. 2023). Moreover, while highlights in the portrait may be distributed sporadically, they do not completely obscure the face. Therefore, this patch-based training approach is beneficial in distinguishing between highlight and non-highlight regions more effectively. This distinction supports more accurate removal of highlight components during the sampling process.

Specifically, for a highlight-free portrait $x \in R^{3\times H\times W}$, we slide-split it into the patch sequences $\{x^1, x^2, ..., x^I\}$, where $x^i \in R^{3\times h\times w}$ and $I = \frac{H\times W}{h\times w}$. Subsequently, we perform forward diffusion on these patches individually. However, the patches within the model are independent of each other at this time. So, we incorporate relative position embeddings (p^i) into the model to enhance its ability to differentiate between various patches.

$$p^i(i, 2j) = \sin(i/10000^{2j/d}), \tag{1}$$

$$p^i(i, 2j+1) = \cos(i/10000^{(2j+1)/d}), \tag{2}$$

where i presents the current patch, d denotes the dimension of the model. $p^i(i, 2j)$ and $p^i(i, 2j+1)$ represent the values of the j-th dimension of the positional embedding at the even and odd positions (Vaswani et al. 2017), respectively. Then, the position embedding of each patch can be defined as:

$$x_i^p = x^i + p^i, i \in \{1, .., I\}. \tag{3}$$

However, the current challenge is that the model only focuses on local patches and fails to capture the global dependencies between them. Global representation is crucial for facial integrity. The absence of global representation may cause the face to appear incoherent in certain areas, or with distorted artifacts. To address this issue, we innovatively propose a Patch-Residual Training Strategy (PRTS), as detailed in Figure 2(b). To be specific, at each diffusion step t, we first encode the full-resolution portrait x into the latent space by an encoder $\varepsilon(\cdot)$. The $\varepsilon(\cdot)$ can map the image from pixel space to latent space, facilitating the extraction of deeper semantic information (Rombach et al. 2022). The encoded x then serves as the residual term added to the forward diffusion process. This residual term acts as a constraint, guiding the forward diffusion process towards a specific data distribution. Note that all inputs are encoded with distinct parameters by $\varepsilon(\cdot)$ before being fed into the diffusion model. This latent encoding process can be expressed as:

$$x', \widetilde{x}_i, y' = \varepsilon(x, x_i^p, y). \tag{4}$$

Finally, the forward diffusion can be defined as:

$$q(\widetilde{x}_{t,i}|\widetilde{x}_{0,i}) = q(\widetilde{x}_{T,i}|\widetilde{x}_{0,i}) \prod_{t=2}^{T} q(\widetilde{x}_{t,i}|\widetilde{x}_{t-1,i}, \widetilde{x}_{0,i}), \tag{5}$$

$$q(\widetilde{x}_{t,i}|\widetilde{x}_{t-1,i},\widetilde{x}_{0,i}) = \mathcal{N}(\widetilde{x}_{t-1,i};\widetilde{\mu}_t(\widetilde{x}_{t,i},\widetilde{x}_{0,i}),y',\mathbf{I}), \quad (6)$$

where $\widetilde{x}_{0,i}$ is the initial latent features of patches i, and y' is the constraint condition. The x' is progressively added as a residual term to the forward process and the mean is denoted in variance as:

$$\widetilde{\mu}_t = \sqrt{\bar{\alpha}_{t-1}}\widetilde{x}_{0,i} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t + x', \quad (7)$$

where $\alpha_t = 1-\beta_t$ (β_t is the noise schedule), $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, and $\epsilon_t \sim \mathcal{N}(0,\mathbf{I})$. Assuming that there is a denoiser f_θ (which will be explained below), and our PHR-DIFF training objective can be defined as follows:

$$\mathcal{L}_{diff} = \| \widetilde{\epsilon_t} - \epsilon_t \|_2^2, \quad (8)$$

where $\widetilde{\epsilon_t} = f_\theta(\widetilde{x}_{t,i}, y, t)$ is the prediction noise. The overall training approach is outlined in Algorithm 1. Notably, when the diffusion is done on each patch, we can decode the p^i of each patch to stitch them together to produce a full-resolution noisy image X_T.

Patch-Aware Highlight Removal Sampling for PHR-DIFF. After training the network, we perform the reverse diffusion process to progressively remove the highlight components in the portrait. The primary challenges in this process are completely removing glaring highlight in the face and restoring the high-fidelity details in the highlight regions, including skin color, texture, and contours, to avoid artifacts. Highlights typically do not cover the entire face, which allows us to find the non-highlight areas as guidance for removing the highlight components. To make the most of the detailed features in these non-highlight areas, a more compact and refined feature encoding approach is required. Thus, we propose a patch-aware highlight removal method to address these challenges effectively.

As demonstrated in Figure 2(a), PHR-DIFF takes the highlight portrait y, the highlight mask m and the pure noise distribution $X_T \in \mathcal{N}(0,\mathbf{I})$ as inputs in the sampling stage, where the m is obtained by the pre-trained detection network (Fu et al. 2024). We leverage the same patchify strategy as in training. The y is split into patch sequences $\{y^1, y^2, ..., y^I\}$ with the same size x^i. Similarly, m is partitioned into $\{m^1, m^2, ..., m^I\}$. The highlight areas (ha) and non-highlight areas (na) in each patch y^i can be calculated by multiplying with m^i as follows:

$$ha = y^i \times m^i, \quad (9)$$

$$na = y^i \times (1-m^i). \quad (10)$$

This patch-aware operation is illustrated in Figure 2(c). Subsequently, we leverage the features in na to guide the removal of highlight in ha. This is accomplished through cross-attention, which facilitates deeper feature interactions:

$$z^i = Atten(Q_{ha}, K_{na}, V_{na}) = Softmax(\frac{Q_{ha}K_{na}^T}{\sqrt{D}})V_{na}, \quad (11)$$

where $Q_{ha} = haW_q$, $K_{na} = naW_k$, $V_{na} = naW_v$, and W_q, W_k, W_v are learnable parameters, and D is the hidden dimension (Zhang et al. 2024). In this way, we collect the attention patch sequences $\{z^1, z^2, ..., z^I\}$, and decoder them by $\tau(\cdot)$:

$$\{z^{1'}, z^{2'}, ..., z^{I'}\} = \tau(\{z^1, z^2, ..., z^I\}). \quad (12)$$

Algorithm 1: Patch-Residual Training for PHR-DIFF

Input: highlight portrait: y, highlight-free portrait: x, patches number: I, latent encoder: $\varepsilon(\cdot)$, and forward diffusion steps: T.

while not converged **do**
  $\{x^1, x^2, ..., x^I\}$ = patchify x into sequences
  $x_i^p = x^i + p^i, i \in \{1, .., I\}$, #$p^i$ is position embedding
  $x', \widetilde{x_i} = \varepsilon(x, x_i^p)$
  t $\sim$ Uniform{1,...,T}
  $\epsilon_t \sim \mathcal{N}(0,\mathbf{I})$
  Perform Gradient descent steps on $\nabla_\theta \mathcal{L}_{diff}(\theta)$.
end while
Output: θ

Algorithm 2: Patch-Aware Sampling for PHR-DIFF

Input: highlight portrait: y, highlight mask: m, diffusion model: f_θ, $X_T \in \mathcal{N}(0,\mathbf{I})$, sampling steps: T, patches number: I, and latent decoder: $\tau(\cdot)$.

for t = I,...,1 **do**
  $\{y^1, y^2, ..., y^I\}$ = patchify y into sequences
  $\{m^1, m^2, ..., m^I\}$ = patchify m into sequences
  $\{x^1, x^2, ..., x^I\}$ = patchify X_T into sequences
  $ha = y^i \times m^i$, $na = y^i \times (1-m^i)$
  z^i = Atten(Q_{ha},K_{na},V_{na})
  $\{z^{1'}, z^{2'}, ..., z^{I'}\} = \tau(\{z^1, z^2, ..., z^I\})$
  for t = T,...,1 **do**
    $x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\cdot f_\theta(x_t, z^{i'}, t)}{\bar{\alpha}_t}) + \sqrt{1-\bar{\alpha}_{t-1}}\cdot f_\theta(x_t, z^{i'}, t)$
  end for
end for
Output: highlight-free portrait: x_t

Finally, the sampling can be defined as the reverse diffusion under some conditions (e.g. specific data distribution: y' or other information). This process starts from a pure noise distribution x_T and reverses to x_0 according to the following formula:

$$p_\theta(x_{0:T}|y) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, z^{i'}, t), \quad (13)$$

$$p_\theta(x_{t-1}|x_t, z^{i'}, t) = \sqrt{\bar{\alpha}_{t-1}}(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\cdot f_\theta(x_t, z^{i'}, t)}{\bar{\alpha}_t}) + \sqrt{1-\bar{\alpha}_{t-1}}\cdot f_\theta(x_t, z^{i'}, t). \quad (14)$$

The overall flow of patch-aware sampling is summarized in Algorithm 2.

Experiments

Experimental Setup

Datasets. We utilize the PFSD dataset (Zheng et al. 2024) for the model training and evaluate our method on several public datasets, including Spec-Face (Muhammad et al. 2020), FFHQ (Karras, Laine, and Aila 2019), and CelebA (Ziwei Liu and Tang 2015), all of which have the resolution with 256×256. For the FFHQ and CelebA, we manu-

ally select 1,150 and 480 portraits with significant specular highlights, respectively. Notably, FFHQ and CelebA do not provide highlight-free ground truth (GT), so only qualitative evaluations are conducted.

Implementation Details. We implement our PHR-DIFF using PyTorch and train it on 6 NVIDIA GeForce RTX 3090 GPUs. The Adam optimizer is employed with parameters as (0.9, 0.999). During the training, we set the diffusion steps T to 1,000, and the noise schedule β_t increases linearly from 0.0001 to 0.02. The model is trained for the 1000 epochs. For the sampling, we use a U-Net architecture similar to (Saharia et al. 2022) as the denoiser f_θ, with 25 sampling steps. The patch size is set to 64×64, resulting in a total of I=4×4 patches. The more detailed architectures and hyperparameter settings can be found in the supplementary.

Evaluation Metrics. We evaluate the performance of our PHR-DIFF using five commonly-used metrics: Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), and DreamSim (Fu et al. 2023b), Structural Similarity Index (SSIM), Multi-Scale Structural Similarity Index Measure (MS-SSIM) and Peak Signal to Noise Ratio (PSNR). In general, higher SSIM, MS-SSIM, and PSNR values, along with lower LPIPS and DreamSim values, indicate better model performance.

Comparisons with State-of-the-arts

We compare our PHR-DIFF with the following five state-of-the-art methods: (Zheng et al. 2024) (Liang et al. 2021; Muhammad et al. 2020) designed for the portrait highlights removal, and (Wu et al. 2023, 2021) proposed for the natural object highlights removal.

Quantitative Comparison. Tables 1 and 2 report the quantitative evaluations of these methods on the PFSD and Spec-Face datasets, respectively. It is noteworthy that our PHR-DIFF is trained solely on the PFSD dataset and evaluated on the Spec-Face dataset without any fine-tuning. The results demonstrate that our method exhibits excellent generalization capabilities, achieving optimal performance on both datasets. The lowest LPIPS and highest SSIM indicate that our removal results are nearly indistinguishable from the GT in terms of perceptual similarity and structural content.

Visual Comparison. Figures 3, 4, and 5 present the visual results on the PFSD, FFHQ, and CelebA datasets, respectively. Comparing the removal results, our approach demonstrates several significant improvements. Firstly, PHR-DIFF effectively removes highlight components with minimal specular residue. Secondly, our method excels in high-fidelity restoration of local details in highlighted regions, providing the most realistic and natural visual experience. Additionally, our approach adapts well to varying illumination conditions without introducing hue deviation to the facial skin. Lastly, PHR-DIFF exhibits strong generalization capabilities, achieving exceptional highlight removal performance even on previously unseen datasets, such as FFHQ, and CelebA. More comparison results can be seen in the supplementary.

User Study. We conduct a user study on the FFHQ dataset to further compare the results. Specifically, we first select 30 test cases from FFHQ, and six corresponding highlight removal results are generated for each input by using these 6 methods above. Subsequently, following (Guo et al. 2021; Cong et al. 2020), two images are then randomly selected from these six to create image pairs, and $30\times \binom{6}{2} = 450$ portrait pairs could be acquired. Then 40 volunteers are recruited to evaluate the quality of these portrait pairs and asked to choose the one that looks more realistic removal result. Finally, we calculate the B-T Score (Bradley and Terry 1952) of each method based on the evaluation results. From Table 3, PHR-DIFF can bring more satisfying results to the subjects with the highest scores.

Methods	LPIPS↓	DreamSim↓	SSIM↑	MS-SSIM↑	PSNR↑
Liang et al. 2021	0.454	0.534	0.859	0.874	22.16
Muhammad et al. 2020	0.327	0.359	0.875	0.887	24.51
Wu et al. 2021	0.187	0.158	0.928	0.933	29.80
Wu et al. 2023	0.211	0.222	0.907	0.918	30.92
Zheng et al. 2024	0.097	0.109	0.945	0.953	33.84
Ours	**0.092**	**0.100**	**0.957**	**0.965**	**35.07**

Table 1: Quantitative comparison on the PFSD dataset, and the best results are bold. Notably, (Wu et al. 2023, 2021) are designed for natural object highlights removal, and (Liang et al. 2021; Muhammad et al. 2020) are specialized for portrait highlights removal.

Methods	LPIPS↓	DreamSim↓	SSIM↑	MS-SSIM↑	PSNR↑
Liang et al. 2021	0.485	0.572	0.833	0.860	21.32
Muhammad et al. 2020	0.348	0.373	0.858	0.899	26.46
Wu et al. 2021	0.202	0.171	0.871	0.890	24.93
Wu et al. 2023	0.186	0.155	0.898	0.903	25.35
Zheng et al. 2024	0.114	0.122	0.908	0.917	31.40
Ours	**0.105**	**0.116**	**0.925**	**0.933**	33.68

Table 2: Quantitative comparison on the Spec-Face dataset (Muhammad et al. 2020).

Methods	B-T Score↑
Liang et al. 2021	0.872
Muhammad et al. 2020	1.026
Wu et al. 2021	1.326
Wu et al. 2023	1.596
Zheng et al. 2024	1.741
Ours	**1.937**

Table 3: User study on FFHQ dataset. A higher B-T score means that the results are more favored to the subjects.

Ablation Study

We perform the ablation study to validate the effectiveness of our PHR-DIFF. First, we investigate the effect of patch sizes on the model. Then we verify the effect of different training and sampling strategies, which includes: baseline without any patchify strategy (Guo et al. 2023a); patchify training without residual; patchify training with residual; patchify sampling without the PAHR module; patch-based sampling with the PAHR module. Here residual denotes the encoded full-resolution image, which gets the facial global representation, used as the residual term to guide training. The PAHR represents the patch-aware highlights removal module with cross-attention. The quantitative results are shown in Table 5 and 4, and the qualitative results of the

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

Figure 3: Qualitative comparison results on PFSD dataset. (a) Input; (b) GT; **(c) Ours**; (d) Zheng et al. 2024; (e) Wu et al. 2023; (f) Wu et al. 2021; (g) Muhammad et al. 2020; (h) Liang et al. 2021.

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

Figure 4: Qualitative comparison results on FFHQ dataset. (a) Input; **(b) Ours**; (c) Zheng et al. 2024; (d) Wu et al. 2023; (e) Wu et al. 2021; (f) Muhammad at al. 2020; (g) Liang at al. 2021.

patch-residual and PAHR are presented in Figure 6.

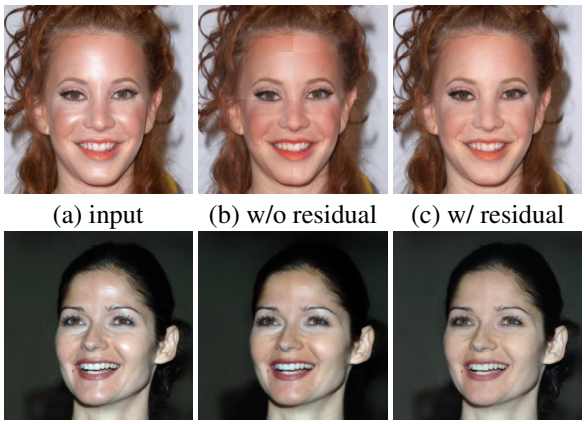
Effect of patch size. The patch size is an essential parameter for our PHR-DIFF, and the model performance for different values of size is reported in Table 4. Patchify helps with performance. However, as the patch size decreases, the performance gain is less and requires more computational burden and sampling time. Considering all, we set the patch size to 64×64.

Effect of training strategy. The second row of Table 5 shows that the patchify strategy can improve the performance compared to directly performing forward diffusion on the full-resolution image. This is mainly because patching helps extract more compact facial features. By extracting finer and more compact features, the model can better understand and recover details in highlight areas such as skin tone, texture, and contours, thereby producing more natural and realistic results.

Effect of patch-residual. The global representation plays an important role in the consistency of the removal result. Without it, visual artifacts and incoherence will occur. The V line in Table 5 and Figure 6 verifies the effectiveness of the residual to introduce global representation. By using this strategy, the final result remains consistency and avoids artifacts.

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

Figure 5: Qualitative comparison results on CelebA dataset. (a) Input; **(b) Ours**; (c) Zheng et al. 2024; (d) Wu et al. 2023; (e) Wu et al. 2021; (f) Muhammad et al. 2020; (g) Liang et al. 2021.



(a) input (b) w/o residual (c) w/ residual

(d) input (e) w/o PAHR (f) w/ PAHR

Figure 6: Effect of our patch-residual and PAHR. Here, w/ residual means that global facial information is introduced as the residual term during training, and w/ PAHR indicates that the patch-aware highlight removal module is used during sampling.

Patch-size	SSIM↑	LPIPS↓	Time (Sec.)↓
128×128	0.939	0.107	1.7
64×64	0.951	0.094	2.5
32×32	0.959	0.089	4.1

Table 4: Effect of the patch size in our PHR-DIFF. Time represents the sampling speed in seconds.

Ablation	Training strategy		Residual	Sampling strategy		PAHR	SSIM↑	LPIPS↓
	full-reso	patchify		full-reso	patchify			
I	√			√			0.919	0.131
II		√		√			0.929	0.123
III	√				√		0.923	0.127
IV		√			√		0.932	0.116
V		√	√	√			0.935	0.110
VI		√	√		√		0.944	0.103
Ours		√	√		√	√	**0.957**	**0.092**

Table 5: Ablation study of our proposed patch-based training and sampling strategy, where full-reso means using the full-resolution portrait to training or sampling.

Effect of sampling strategy. The third row in Table 5 shows the results of incorporating patch-wise sampling based on patch-residual training. It can be observed that the patch-wise sampling approach also contributes to improved model performance. Patch-wise sampling allows local optimization of each image patch, which can more accurately control the highlight components removal in portrait.

Effect of PAHR module. The proposed patch-aware highlight removal module, shown in the last line of Table 5 and Figure 6, effectively utilizes features from highlight-free regions in portraits. It enhances the deep interaction between features of highlight and non-highlight through cross-attention, resulting in cleaner highlight removal and more natural restoration of local details in the portrait.

Conclusion and Discussion

In this work, we have developed a patch-based diffusion model tailored for portrait highlight removal, termed PHR-DIFF, to rectify the impact of specular highlights on portrait quality. PHR-DIFF employs a patch-based training and sampling strategy, which not only enables the extraction of more compact facial features but also effectively utilizes non-highlight regions to guide the removal and restoration of highlight areas. Extensive experiments validated the superiority of our method.

PHR-DIFF still has several limitations. Firstly, in scenes with strong highlight, the high-frequency information in the highlight regions is completely obscured, making it challenging to achieve clean removal. Secondly, when the original facial skin tone is too dark or too bright, the limited features of the surrounding non-highlight regions provide insufficient guidance, resulting in suboptimal outcomes.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 61972298 and No. 62372336).

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv preprint arXiv:2302.08113*.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Cong, W.; Zhang, J.; Niu, L.; Liu, L.; Ling, Z.; Li, W.; and Zhang, L. 2020. DoveNet: Deep Image Harmonization via Domain Verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.

Ding, Z.; Zhang, M.; Wu, J.; and Tu, Z. 2023. Patched denoising diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Feng, W.; Cheng, X.; Sun, J.; Xiong, Z.; and Zhai, Z. 2023. Specular highlight removal and depth estimation based on polarization characteristics of light field. *Optics Communications*, 537: 129467.

Fu, G.; Zhang, Q.; Zhu, L.; Li, P.; and Xiao, C. 2021. A multi-task network for joint specular highlight detection and removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7752–7761.

Fu, G.; Zhang, Q.; Zhu, L.; Lin, Q.; Wang, Y.; Fan, S.; and Xiao, C. 2024. Towards high-resolution specular highlight detection. *International Journal of Computer Vision*, 132(1): 95–117.

Fu, G.; Zhang, Q.; Zhu, L.; Xiao, C.; and Li, P. 2023a. Towards High-Quality Specular Highlight Removal by Leveraging Large-Scale Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12857–12865.

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023b. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.

Guo; Xiaojie; Cao, X.; and Ma, Y. 2014. Robust separation of reflection from multiple images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2187–2194.

Guo, L.; Wang, C.; Yang, W.; Huang, S.; Wang, Y.; Pfister, H.; and Wen, B. 2023a. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14049–14058.

Guo, L.; Wang, C.; Yang, W.; Wang, Y.; and Wen, B. 2023b. Boundary-Aware Divide and Conquer: A Diffusion-based Solution for Unsupervised Shadow Removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13045–13054.

Guo, Z.; Zheng, H.; Jiang, Y.; Gu, Z.; and Zheng, B. 2021. Intrinsic Image Harmonization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 16367–16376.

Ho; Jonathan; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, H.; Jin, H.; Hadap, S.; and Kweon, I. 2013. Specular Reflection Separation Using Dark Channel Prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li; Chen; Zhou, K.; and Lin, S. 2015a. Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 4621–4629.

Li, C.; Lin, S.; Zhou, K.; and Ikeuchi, K. 2017. Specular highlight removal in facial images. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 3107–3116.

Li, F.; Tian, J.; Tang, Y.; and Wang, Y. 2015b. An image highlights removal method with polarization principle. In *2015 3rd International Conference on Machinery, Materials and Information Technology Applications*, 402–407. Atlantis Press.

Liang, B.; Weng, D.; Tu, Z.; Luo, L.; and Hao, J. 2021. Research on face specular removal and intrinsic decomposition based on polarization characteristics. *Optics Express*, 29(20): 32256–32270.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Mallick, S. P.; Zickler, T.; Belhumeur, P. N.; and Kriegman, D. J. 2006. Specularity removal in images and videos: A PDE approach. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, 550–563. Springer.

Muhammad, S.; Dailey, M. N.; Farooq, M.; Majeed, M. F.; and Ekpanyapong, M. 2020. Spec-Net and Spec-CGAN: Deep learning models for specularity removal from faces. *Image and Vision Computing*, 93: 103823.

Ozan, O.; and Robert, L. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–12.

Ponglertnapakorn; Puntawat; Tritrong, N.; and Suwajanakorn, S. 2023. DiFaReli: Diffusion face relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22646–22657.

Quinn, N. A.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Sauer, A.; Boesel, F.; Dockhorn, T.; Blattmann, A.; Esser, P.; and Rombach, R. 2024. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Son; Minjung; Lee, Y.; and Chang, H. S. 2020. Toward specular removal from natural images based on statistical reflection models. *IEEE Transactions on Image Processing*, 29: 4204–4218.

Song; Jiaming; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Su, T.; Zhou, Y.; Yu, Y.; and Du, S. 2022. Highlight Removal of Multi-View Facial Images. *Sensors*, 22(17): 6656.

Tan, R.; and Ikeuchi, K. 2005. Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2): 178–193.

Todd, J. T.; Norman, J. F.; and Mingolla, E. 2004. Lightness constancy in the presence of specular highlights. *Psychological Science*, 15(1): 33–39.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, Z.; Jiang, Y.; Zheng, H.; Wang, P.; He, P.; Wang, Z.; Chen, W.; Zhou, M.; et al. 2024. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36.

Wang, Z.; Lu, M.; Xu, F.; and Cao, X. 2021. In-the-Wild Facial Highlight Removal via Generative Adversarial Networks. In *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part I 1*, 311–322. Springer.

Wu, Z.; Guo, J.; Zhuang, C.; Xiao, J.; Yan, D.-M.; and Zhang, X. 2023. Joint specular highlight detection and removal in single images via Unet-Transformer. *Computational Visual Media*, 9(1): 141–154.

Wu, Z.; Zhuang, C.; Shi, J.; Guo, J.; Xiao, J.; Zhang, X.; and Yan, D.-M. 2021. Single-image specular highlight removal via real-world dataset construction. *IEEE Transactions on Multimedia*, 24: 3782–3793.

Wu, Z.; Zhuang, C.; Shi, J.; Xiao, J.; and Guo, J. 2020. Deep specular highlight removal for single real-world image. In *SIGGRAPH Asia 2020 Posters*, 1–2.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Ye, T.; Chen, S.; Chai, W.; Xing, Z.; Qin, J.; Lin, G.; and Zhu, L. 2024. Learning Diffusion Texture Priors for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2524–2534.

Yi; Renjiao; Tan, P.; and Lin, S. 2020. Leveraging multiview image sets for unsupervised intrinsic image decomposition and highlight separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12685–12692.

Yu, Z.; Zhou, J.; Bao, Z.; Fu, G.; He, W.; Liang, C.; and Xiao, C. 2024. CFDiffusion: Controllable Foreground Relighting in Image Compositing via Diffusion Model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3647–3656.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Huang, J.; Zhou, Q.; Wang, Z.; Wang, F.; Luo, J.; and Yan, J. 2024. Continuous-Multiple Image Outpainting in One-Step via Positional Query and A Diffusion-based Approach. *arXiv preprint arXiv:2401.15652*.

Zheng, H.; Xu, W.; Wang, Z.; Lu, X.; and Xiao, C. 2024. Facial Highlight Removal with Cross-Context Attention and Texture Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Zhou, J.; Yu, Z.; Bao, Z.; Fu, G.; He, W.; Liang, C.; and Xiao, C. 2024. Foreground Harmonization and Shadow Generation for Composite Image. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8267–8276.

Ziwei Liu, X. W., Ping Luo; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.