Diff-SemiER: Transparency-Aware Adaptive Fusion Diffusion Model with Generative Prior for Semi-Transparent Eyeglasses Removal

Supplementary Material

In the supplementary material, we first present additional examples synthesized by our proposed semi-transparent eyeglasses generation method to further demonstrate its quality and diversity. We then provide detailed experimental configurations to facilitate reproducibility. Moreover, we include extended visual results on both the synthetic and real-world datasets, along with additional comparisons against existing methods, comprehensively showcasing the robustness and superiority of Diff-SemiER under varying occlusion levels and lens transparencies. Finally, we provide several failure cases and analyze their underlying causes to highlight the current limitations of our model.

1. Additional Synthesized Eyeglasses Samples

Fig. 1 showcases additional samples produced by our transmittance-based semi-transparent eyeglasses synthesis method. The results illustrate the method's ability to generate realistic lens appearances across a wide range of transparency levels while preserving underlying facial structures. These visual examples further highlight the quality and diversity of the synthesized data used for training and evaluation.

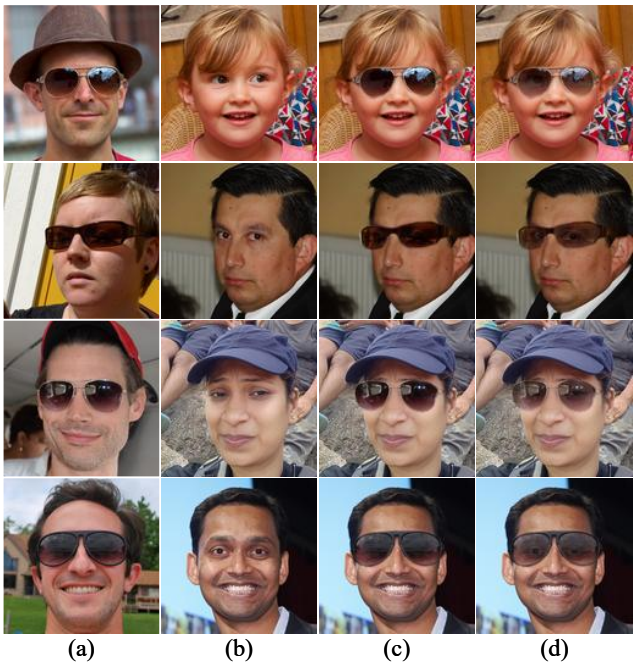

Figure 1. Additional visual results of the proposed semi-transparent eyeglass synthesis pipeline. (a) Real faces with opaque sunglasses. (b) Eyeglass-free faces. (c) Synthesized faces with opaque eyeglasses. (d) Synthesized faces with semi-transparent eyeglasses.

2. Details of Experimental Configurations

Implementation Details. The proposed method is implemented using Python and PyTorch. Each module is trained independently for 200 epochs. We use the Adam optimizer with momentum as (0.9, 0.999) to optimize the TPM, GPDM, and TAFDM, with an initial learning rate set to 5×10^{-5}. Following [9], we use the Kaiming initialization technique [1] to initialize the weights of the proposed model and use 0.9999 Exponential Moving Average (EMA) for all our experiments. For the GPDM and TAFDM, we adopt a U-Net architecture similar to the denoiser in [9]. The noise schedule β_t increases linearly from 0.0001 to 0.02. During training, the batch size and image size are set to 8 and 128 $\times$ 128, respectively. The diffusion step T is set as 1000 for training and 100 for inference. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

Datasets. Our experiments primarily use three public facial datasets: FFHQ, CelebA, and CelebA-HQ. The FFHQ dataset contains 70,000 high-quality facial images, covering a wide range of ages, genders, poses, and diverse real-world accessories (including various types of eyeglasses). It serves as a common benchmark for face generation and editing tasks. The CelebA and CelebA-HQ datasets provide large-scale facial images and their high-resolution counterparts, respectively, annotated with diverse attribute labels, and widely used in facial attribute editing and generation research.

Based on the semi-transparent eyeglasses synthesis method described in Section 3.1, we generated 25,000 paired training samples from the FFHQ dataset. Among these, 10,000 pairs were used to train the GPDM, and 15,000 pairs were used to train the TAFDM. For the testing phase, we synthesized an additional 1,000 paired samples from the FFHQ dataset to construct the FFHQ-Test synthetic test set, which is used to evaluate the model's generalization ability under varying data distributions. Furthermore, we selected a subset of real-world facial images with semi-transparent eyeglasses from the aforementioned datasets to construct a real-world test set, aiming to verify the applicability and robustness of our method in practical scenarios. It is worth noting that in real-world datasets, strictly paired eyeglass-free reference images correspond-

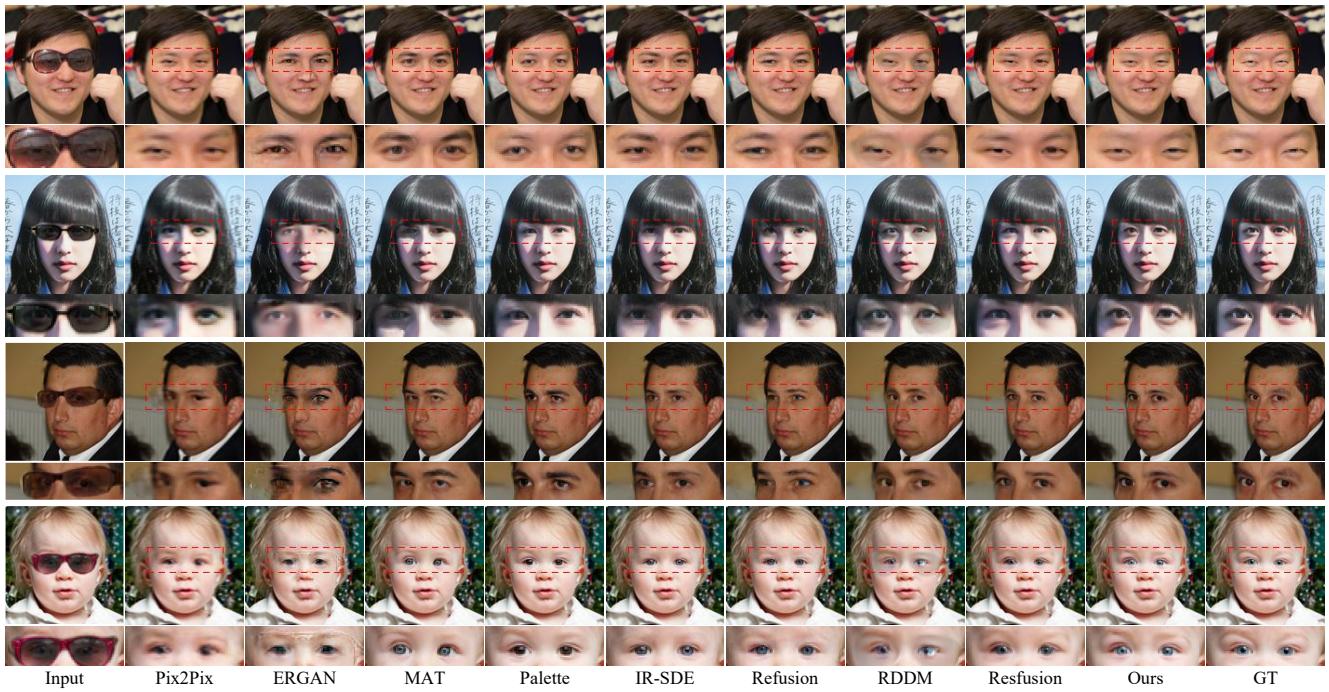

Figure 2. Qualitative comparison on the FFHQ-Test synthetic dataset. Our Diff-SemiER restores more natural and detailed eye regions. Please **zoom in** for a better view of the details.

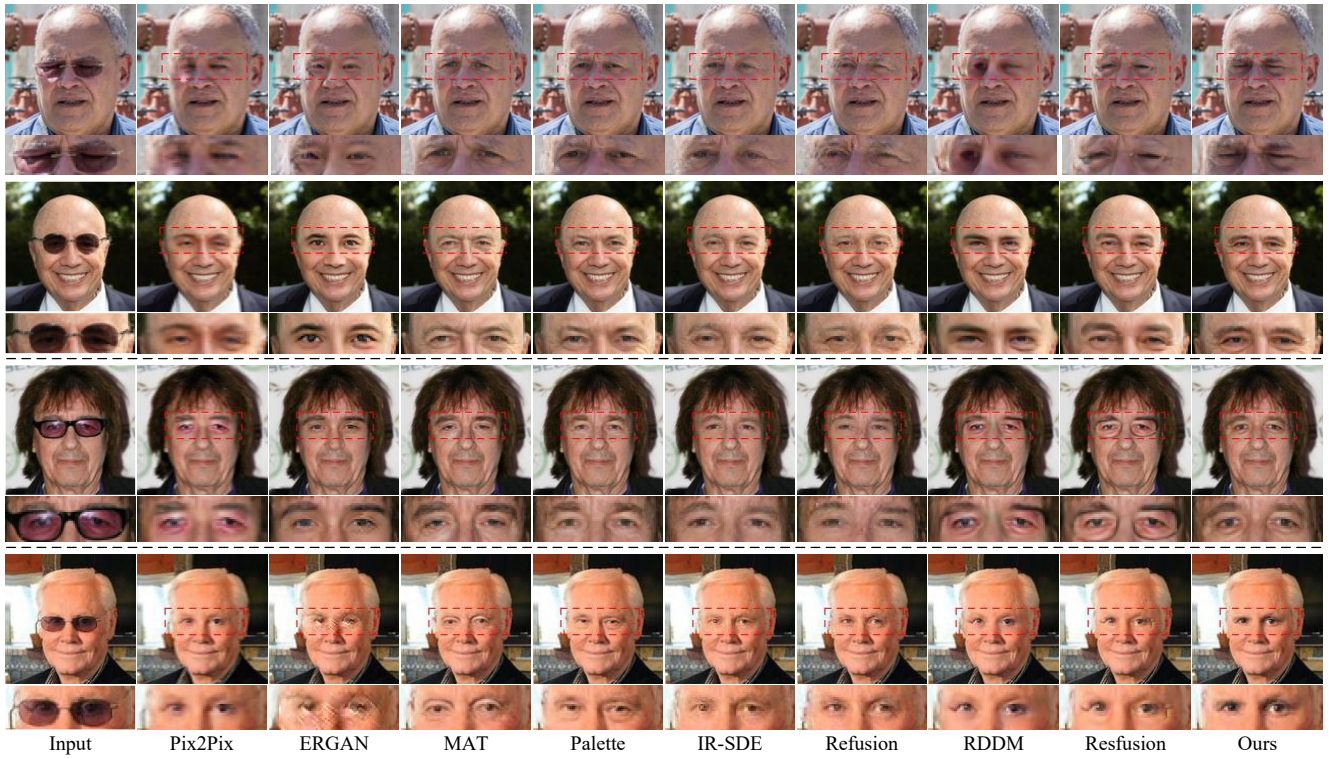

Figure 3. Qualitative comparison on the three real-world datasets: FFHQ, CelebA-HQ, and CelebA (from top to bottom, respectively). Please **zoom in** for a better view of the details.

ing to the test samples are not available.

3. Additional Comparison with Existing Methods

We provide additional qualitative results on our synthetic FFHQ-Test dataset in Fig. 2, comparing our method with Pix2Pix [3], ERGAN [2], MAT [4], Palette [8], IR-SDE [6], Refusion [7], RDDM [5], and Resfusion [10]. Similarly, we present more results on real-world facial images from FFHQ, CelebA, and CelebA-HQ in Fig. 3 with comparisons against the same set of baseline methods.

3.1. More Visualization Results on the Synthetic Dataset

From Fig. 2, we can observe that although competing methods can remove eyeglasses, they often introduce boundary artifacts and blurriness (e.g., Pix2Pix, ERGAN), or suffer from identity drift due to neglecting visible cues (e.g., MAT, Palette). Furthermore, residual-based methods (e.g., RDDM) tend to generate distortions under strong occlusion. In contrast, our method effectively removes semi-transparent lenses and restores clear, identity-consistent eye details.

3.2. More Visualization Results on Real-World Datasets

From Fig. 3, we can see that in real-world scenarios without paired ground truth, competing methods frequently hallucinate mismatched eye structures or alter the original gaze direction (e.g., changing eye shape). In contrast, our method leverages visible semi-transparent cues to restore realistic eye details such as pupils and eyelashes, maintaining high identity consistency and naturalness.

References

[1] 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. 1

[2] Bingwen Hu, Zhedong Zheng, Ping Liu, Wankou Yang, and Mingwu Ren. Unsupervised eyeglasses removal in the wild. *IEEE transactions on cybernetics*, 51(9):4373–4385, 2020. 3

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[4] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 3

[5] Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2773–2783, 2024. 3

[6] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023. 3

[7] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691, 2023. 3

[8] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 3

[9] 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. 1

[10] Zhenning Shi, Chen Xu, Changsheng Dong, Bin Pan, Along He, Tao Li, Huazhu Fu, et al. Resfusion: Denoising diffusion probabilistic models for image restoration based on prior residual noise. *Advances in Neural Information Processing Systems*, 37:130664–130693, 2024. 3