Supplementary material for: Realistic and Efficient Face Swapping: A Unified Approach with Diffusion Models

Sanoojan Baliah[1] Qinliang Lin[2] Shengcai Liao[3] Xiaodan Liang[1] Muhammad Haris Khan[1]
[1]MBZUAI, UAE [2]Shenzhen University, China [3]United Arab Emirates University, UAE
{sanoojan.baliah, muhammad.haris}@mbzuai.ac.ae

0.1. Diffusion Process Visualization

Fig 1 shows the source, target, and the decoded output images through the diffusion process. Here, we show the denoising through 75 DDIM steps. The $z_{1000} \cdots z_0$ images in Fig 1 correspond to 0^{th}, 30^{th}, 50^{th}, 65^{th}, 70^{th}, 73^{th}, and 75^{th} DDIM step respectively. We observe initial steps recover the basic structure of the swapped image while a few later (65^{th} to 75^{th}) steps refine the image.

0.2. Effect on Number of Steps in DDIM

Existing diffusion-based works on face-swapping use time-consuming denoising steps. DiffSwap [7] uses 200 steps with masked fusion at inference. DiffFace [3] uses 75 steps which consumes a huge amount of time (approximately 9 hours to perform 1000 swaps) due to gradient computation in their inference strategy. We use 50 steps to compare the inference time in which we showed our method performs a magnitude faster inference than DiffFace [3] and roughly twice faster than DiffSwap [7]. Further, all the results we show in the main manuscript use 50 steps. However, to analyze the effect of the number of DDIM steps with our algorithm, we show further analysis of qualitative images with varying numbers of DDIM steps in Fig 2.

In contrast to existing methods that rely on complex inference processes for face-swapping, often failing to produce satisfactory results and fail to transfer identity features well with minimal denoising steps, our approach stands out by achieving superior swapping outcomes even with as few as 5 steps (see Fig. 3), which significantly reduces the computational overhead, slashing the inference time to approximately one-tenth of the duration required for 50 steps.

0.3. Additional Implementation Details

Building upon the implementation details outlined in the main manuscript, we provide additional specifications for clarity. We adopt a pre-trained stable diffusion checkpoint, akin to the framework introduced in [6], with a modification involving 9 channels. We use AdamW optimizer with learning rate $1e-5$ and other default parameters. Latent size is 64×64. The condition feature dimension D is 768. In condition generation, the CLIP weight w_{clip}, ID feature weight w_{id}, and the landmark feature weight w_{lm} are 1.0, 10.0, and 0.05 respectively. The number of DDIM steps in our training pipeline $N = 4$. The output image resolution is 512×512.

0.4. Additional Information on Face shape Augmentation

To facilitate face shape augmentation, an image elastic deformation approach [2, 4] based on Thin Plate Spline (TPS) transformation [1] was employed. Specifically, we first generate a 2D grid of points of the same size as the face mask. Then, we set up a set of control net points O on the grid. Next, we add random noise δ to control net points O and obtain P. The intensity of the noise is controlled by a scaling factor s to enable precise modulation. By utilizing the two sets of control points, we can obtain an interpolation function that acts on the entire mask, allowing us to achieve our mask augmentation while ensuring coherence and consistency in subsequent transformations. This is crucial for enhancing diversity and realism in augmented face shapes, providing smooth and continuous deformation while preserving structural integrity. We use this face shape augmentation to get the inpaint image in our pipeline. We use a random scale s sampled uniformly from the range 0.5 to 1.

0.5. Additional Qualitative Results

We provide more qualitative comparison for Face Swapping on CelebA dataset (Fig 4) and FFHQ dataset (Fig 5). In the examples, we observe, our method produces smoother boundaries and photo-realistic images. Unlike other works which suffer in merging the swapped image and the target's hair and background, which often results in a visible merging boundary our method seamlessly blends the swapped image as there is no separate step of merging. Moreover, other works produce a lot of artifacts, especially in challenging situations such as extreme pose variations (E.g., last row of Fig 4), and occlusions or accessories in source (E.g.,

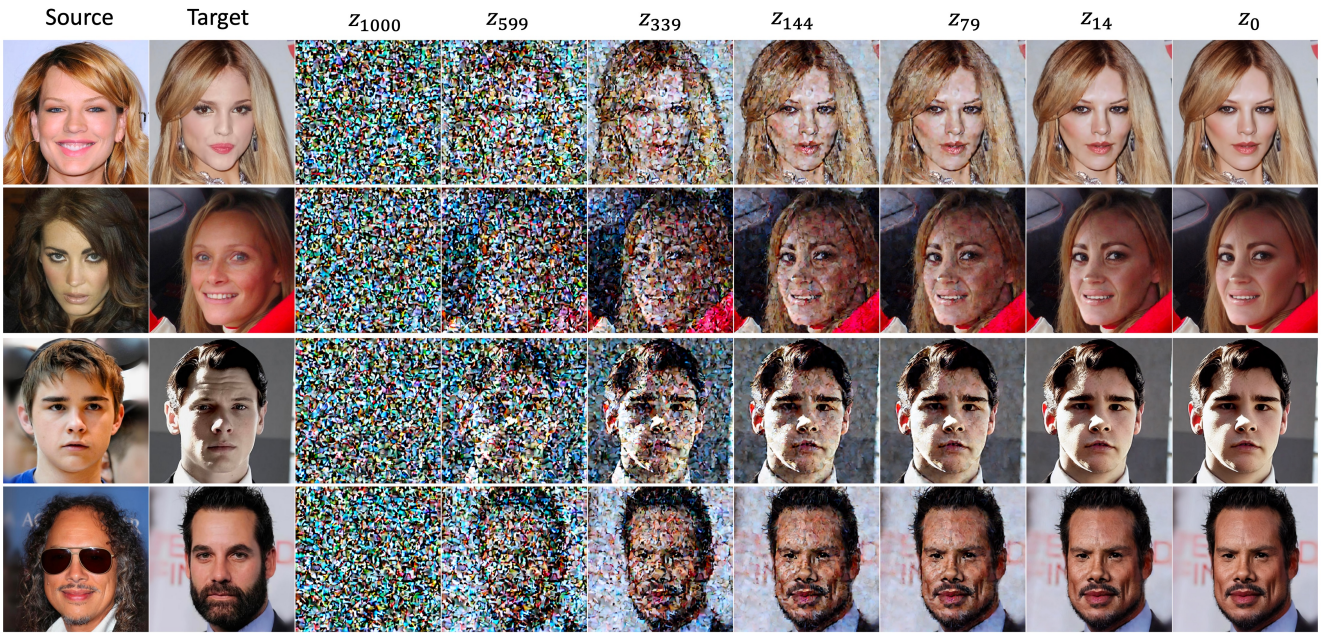

Figure 1. Denoising process visualization. The images shows the decoded output of noisy latents ($\mathcal{D}(z_t)$) through DDIM process.

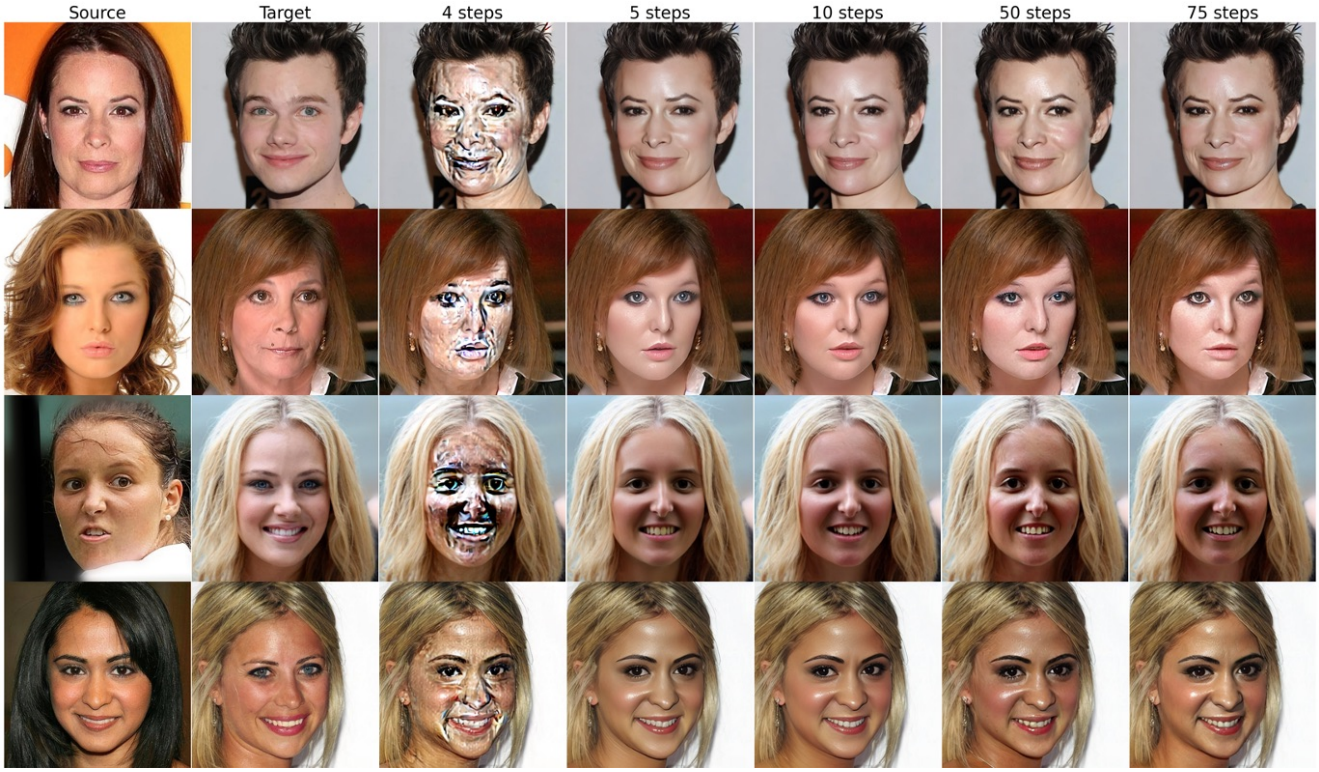

Figure 2. Comparison for total number of denoising steps using DDIM with our model.

third last row of Fig 5).

Further, we provide additional head-swapping qualitative images in Fig 6. Despite challenging masks, our approach is capable of producing realistic head swaps while preserving the target pose and expression.

Figure 3. Comparison of the effect of total number of denoising steps using DDIM with other approaches. While both DiffSwap [7] and DiffFace [3] are failing to transfer identity and making artifacts, our method produces superior swapped images even with 5 steps

0.6. Societal Impact

With the advancements in deep learning, creating swapped face has become easier and social media platforms have made it easier for them to spread rapidly. Everything has two sides. If facial swapping technology is used for the advancement of productivity, such as in movie scenes, it can greatly enhance productivity. However, if this technology is exploited by malicious individuals, face swapping may pose a significant threat to society. We are committed to developing powerful face swapping technologies that have a beneficial impact on society. The purpose of our research is also to promote the healthy development of this technology. Furthermore, we control the generation of vulnerable images in our method's safety check via Stable Diffusion [5].

Figure 4. Qualitative comparison on CelebA dataset. Better viewed in Zoom

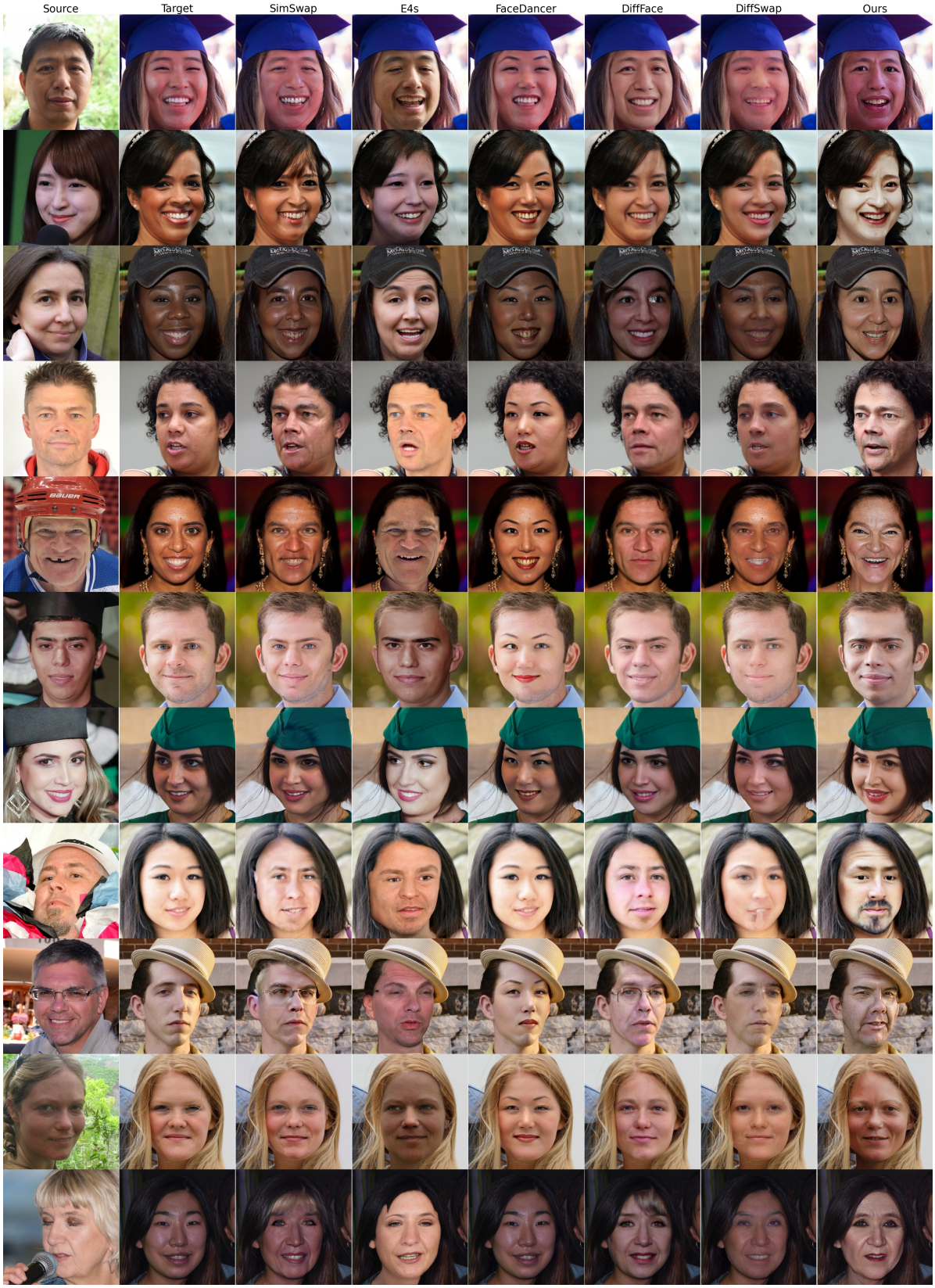

Figure 5. Qualitative comparison on FFHQ dataset. Better viewed in Zoom

Figure 6. Additional Head Swap outcomes. Better viewed in Zoom

References

[1] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 1

[2] Xilin He, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Feng Liu, and Linlin Shen. Shift from texture-bias to shape-bias: Edge deformation-based augmentation for robust object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1526–1535, 2023. 1

[3] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022. 1, 3

[4] Qinliang Lin, Cheng Luo, Zenghao Niu, Xilin He, Weicheng Xie, Yuanbo Hou, Linlin Shen, and Siyang Song. Boosting adversarial transferability across model genus by deformation-constrained warping. *arXiv preprint arXiv:2402.03951*, 2024. 1

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[6] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 1

[7] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023. 1, 3