Generative Unlearning for Any Identity

Supplementary Material

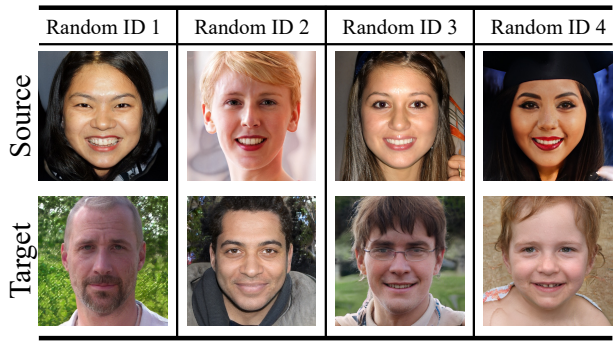

Figure 1. Illustration of diverse target images from randomly generated source images.

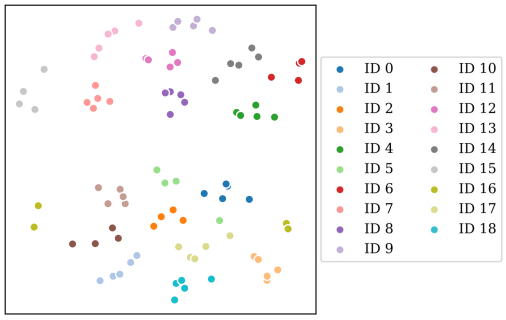

Figure 2. The relationship between the images and their identities in the latent space with t-SNE [6]. Points of the same color denote the same identity. We used 5 images per identity from CelebAHQ dataset.

A. Additional Experiments

A.1. Target Images from Diverse Source Images

In Figure 1, we visualized target images corresponding to the diverse source images. In this experiment, we used randomly sampled w_u to find the corresponding w_t. Instead of the average latent code, by setting an extrapolated latent code as a target, we can obtain the effective and diverse target images for unlearning procedure.

A.2. Distribution of Identities within CelebAHQ

In designing GUIDE, we assume that images sharing the same identity tend to cluster together. Consequently, considering the proximity of latent codes aids in a more comprehensive erasure of identity. Figure 2 illustrates the relationship between images and their respective identities, utilizing 5 images per identity. Our observation reveals a close grouping of images from the same identity in the latent space. This finding aligns with the effectiveness of the $\mathcal{L}_{adj}$ proposed in Section 3.3 of our main paper.

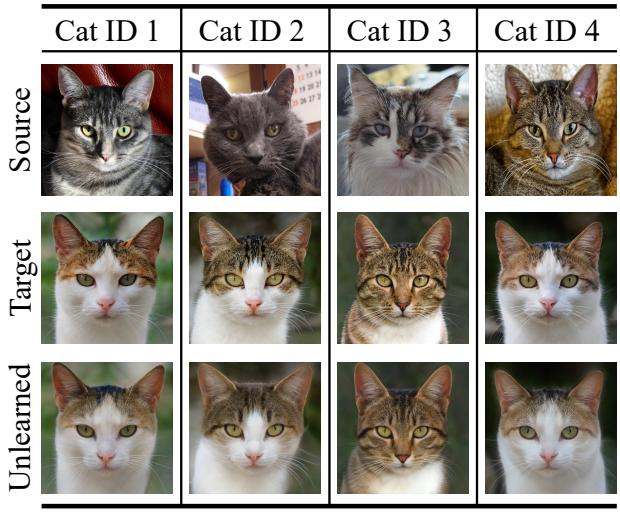

Figure 3. Qualitative results of generative identity unlearning on AFHQv2-Cat dataset.

	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
AFHQv2-Cat	5.93 ± 1.03	3.44 ± 1.68

Table 1. Quantitative results of generative identity unlearning on AFHQv2-Cat dataset. The existing ID metrics are designed for the human face, and there are no adequate metrics for cat. For this reason, we only represent about FID_{pre} and $\Delta\text{FID}_{\text{real}}$ in this experiments.

A.3. Generative Identity Unlearning in AFHQv2

In this section, we validated GUIDE in a different dataset - AFHQv2-Cat [2]. We used the generator architecture [1] and the GAN inversion network [7] pre-trained on AFHQv2-Cat. The pre-trained weights are publicly available at their official implementations. Since the identity loss [3] used in our main experiment were designed to capture the dissimilarities between identities in human faces, we only adopted to use the reconstruction loss and the perceptual loss [8] in this experiment. The qualitative results are shown on Figure 3. We could show the effectiveness of GUIDE in a different domain - faces of cats. In Table 1, we additionally show that GUIDE can preserve performance of pre-trained model on AFHQv2-Cat.

A.4. Target Images from Different d

In addition to Section 4.3 in our main paper, titled "Effect of d in Determination of w_t", this section presents further experiments involving a variety of target images. We visu-

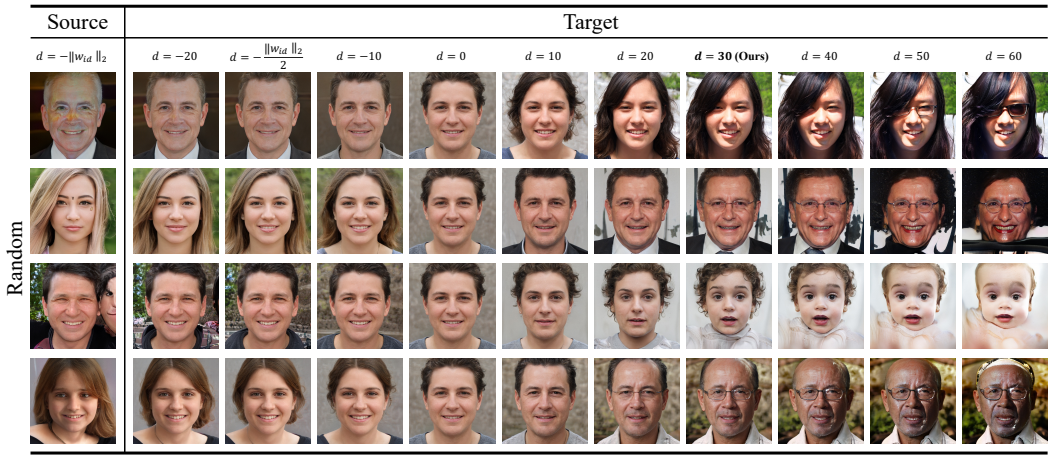

Figure 4. Illustration of target images from source images with different d in *Random* scenario.

Figure 5. Illustration of target images from source images with different d in *In-domain* (FFHQ) and *Out-of-domain* (CelebAHQ) scenario.

alized target images derived from a given source image at multiple d values. In Figure 4, we utilized various d values to sample the corresponding target image in the *Random* scenario, while Figure 5 is for the *In-domain* and *Out-of-domain* scenarios. Our results illustrate that adjusting d allows us to obtain diverse target images. However, as mentioned in our main paper, target images derived from interpolated latent code, where d is less than 0, exhibit similarity to the given source image. Conversely, target images with $d \geq 50$ tend to be corrupted. Therefore, our choice of $d = 30$ appears to strike a visually balanced representation for the target image.

Figure 6. Additional qualitative results with CelebAHQ dataset.

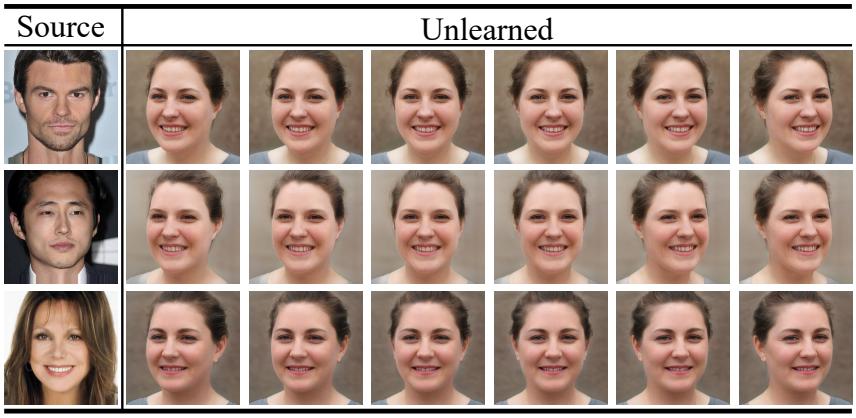

Figure 7. Generated images given the source image with different camera poses.

A.5. Additional Qualitative Results

In this section, we presented additional qualitative results. In contrast to our main paper, we utilized 10 images per identity, and the results are illustrated in Figure 6. The findings emphasize once again that GUIDE successfully erases the identity not only in the given source image but also in other images with the same identity.

A.6. Multi-View Synthesized Images

In this section, we visualized the unlearned images from continuous camera poses. We conducted this experiment within *Out-of-domain* scenario. As shown in Figure 7, our unlearning process successfully erased the source identity across multiple poses.

Method	In-Domain (FFHQ)			Out-of-Domain (CelebAHQ)		
	ID	FID_{pre}	$\Delta\text{FID}_{\text{real}}$	ID	FID_{pre}	$\Delta\text{FID}_{\text{real}}$
Baseline	0.14	8.60	5.97	0.05	6.75	4.32
Ours	**0.06**	**6.14**	**4.35**	**0.01**	**6.07**	**4.25**

Table 2. Quantitative results of GUIDE-SG2 (Ours) and the baseline in the generative identity unlearning task, tested in a single-image setting using one image per identity. Due to space limit, we presented the corresponding standard deviations in the Supplementary materials.

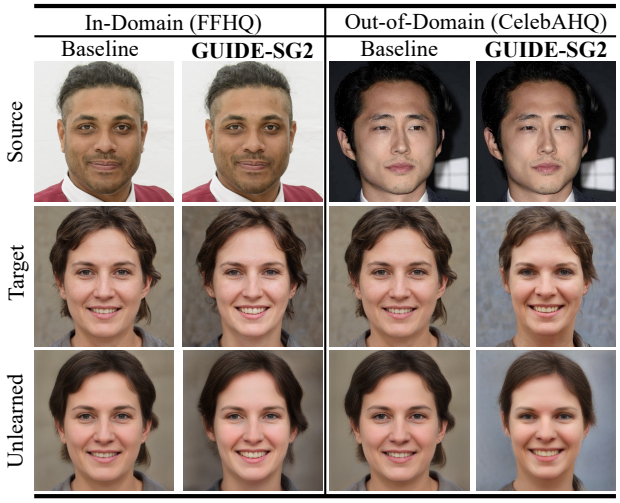

Figure 8. Qualitative results of GUIDE-SG2 and the baseline. For the given source image each (the first row), GUIDE-SG2 and the baseline tried to erase the identity in the pre-trained generator. The result are shown on the second row. Images in the third row are the target image in our unlearning process.

A.7. Generative Unlearning in StyleGAN2

In addition to our primary experiments employing a 3D generative adversarial network as the generator architecture, we observed the effectiveness of our framework in unlearning identity in 2D generative adversarial networks. In this section, we utilized the widely-used StyleGAN2 [4] as the backbone architecture and pSp [5] as a GAN inversion network for latent code extraction from images. Both the backbone and the GAN inversion network were pre-trained on FFHQ [4]. We refer to our framework built on top of StyleGAN2 as GUIDE-SG2.

In GUIDE-SG2, we employed images from the StyleGAN2 generator for calculating loss, instead of tri-plane feature maps. We present results of GUIDE-SG2 qualitatively in Figure 8 and quantitatively in Table 2. Both results demonstrated GUIDE-SG2 successfully erased the given identity in a 2D GAN architecture with minimal impact on the performance of the pre-trained generator.

N_a	ID (↓)	$\text{ID}_{\text{others}}$ (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
1	0.046 ± 0.054	0.191 ± 0.077	8.001 ± 1.992	3.515 ± 1.153
2	**0.030 ± 0.051**	**0.174 ± 0.081**	7.882 ± 1.958	3.442 ± 1.104
4	0.034 ± 0.055	0.184 ± 0.077	**7.668 ± 1.807**	**3.340 ± 1.028**

Table 3. Ablation study to figure out the optimal N_a. To find optimal, we performed the analysis with different N_a. As can be seen, when N_a is 2, GUIDE erase the identity most effectively, and the performance of the pre-trained model can be preserved. We used CelebAHQ dataset in this experiment.

N_g	ID (↓)	$\text{ID}_{\text{others}}$ (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
1	0.065 ± 0.067	0.180 ± 0.080	7.875 ± 2.017	3.491 ± 1.118
2	**0.030 ± 0.051**	**0.174 ± 0.081**	7.882 ± 1.958	3.442 ± 1.104
4	0.031 ± 0.055	0.181 ± 0.075	**7.705 ± 1.868**	**3.359 ± 1.079**

Table 4. Ablation study to figure out the optimal N_g. To find optimal, we performed the analysis with different N_g. As can be seen, when N_g increase, GUIDE can preserve the performance of pre-trained model more effectively. When $N_g = 2$, GUIDE have achieved a balanced performance in our metric. We used CelebAHQ dataset in this experiment.

λ_{L2}	ID (↓)	$\text{ID}_{\text{others}}$ (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
10^{-3}	0.089 ± 0.060	0.269 ± 0.086	**4.815 ± 1.000**	**1.454 ± 0.294**
$\mathbf{10^{-2}}$	**0.030 ± 0.051**	0.174 ± 0.081	7.882 ± 1.958	3.442 ± 1.104
10^{-1}	0.032 ± 0.053	**0.159 ± 0.073**	13.308 ± 2.989	7.783 ± 1.808
1	0.036 ± 0.052	0.161 ± 0.073	15.034 ± 3.079	9.198 ± 1.858

Table 5. Ablation study to figure our the optimal λ_{L2}. We compared the performance among different λ_{L2} in CelebAHQ dataset.

B. Additional Ablation Study

B.1. Number of Latent Codes in Loss Functions

In the computation of $\mathcal{L}_{adj}$ and $\mathcal{L}_{global}$, as outlined in Section 3.3 of our main paper, we incorporated N_a and N_g latent codes, respectively. In this section, we investigate the influence of varying N_a and N_g. Due to an out-of-memory issue in VRAM, these experiments were conducted on an NVIDIA A6000 GPU. Table 3 presents the results of varying N_a in $\mathcal{L}_{adj}$ while keeping N_g fixed at 2. Our findings indicate that using $N_a = 2$ yields the best performance in erasing the given identity among different values of N_a, while maintaining comparable performance in preserving generation quality. Conversely, in Table 4, we varied N_g in $\mathcal{L}_{global}$ while keeping N_a fixed at 2. Results show that using $N_g = 2$ achieves a balanced performance between erasing the given identity and preserving generation performance. Importantly, all cases experimented upon outperformed the baseline in generative identity unlearning task.

λ_{id}	ID (↓)	$\text{ID}_{\text{others}}$ (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
10^{-2}	0.033 ± 0.053	0.177 ± 0.080	**7.879 ± 1.943**	3.460 ± 1.093
$\mathbf{10^{-1}}$	**0.030 ± 0.051**	**0.174 ± 0.081**	7.882 ± 1.958	**3.442 ± 1.104**
1	0.105 ± 0.053	0.244 ± 0.081	7.920 ± 1.781	3.534 ± 0.953

Table 6. Ablation study to figure our the optimal λ_{id}. We compared the performance among different λ_{id} in CelebAHQ dataset.

B.2. Scaling Factors of Loss Functions

In $\mathcal{L}_{local}$ and $\mathcal{L}_{adj}$, as proposed in Section 3.3 of our main paper, we set the scaling factors as $\lambda_{L2} = 10^{-2}$ and $\lambda_{id} = 10^{-1}$. In this section, we conducted ablation studies to determine the effective scaling factors.

In Table 5, we varied λ_{L2} while fixing the other scaling factors as default. For small λ_{L2}, the generator architecture could not successfully erase the given identity. On the other hand, for larger λ_{L2}, the generator architecture lost generation performance significantly. Based on these observations, we decided to use $\lambda_{L2} = 10^{-2}$ for balanced performance. In Table 6, we varied λ_{id} while fixing the other scaling factors as default. Our findings indicate that using $\lambda_{id} = 10^{-1}$ is the most effective.

C. Additional Implementation Details

Besides the Section 4.1 in our main paper, in this section, we additionally describe the implementation details that are omitted in the main paper due to space limit. We ran GUIDE and the baseline using a single NVIDIA A5000 GPU. Erasing the given source identity using GUIDE takes about 20 minutes. We utilized only a single image to represent a certain identity; GUIDE underwent 1,000 iterations throughout our experiments.

References

[1] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 1

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[5] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2287–2296, 2021. 4

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 1

[7] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2437–2447, 2023. 1

[8] 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 (CVPR)*, pages 586–595, 2018. 1