Generative Unlearning for Any Identity

Juwon Seo[1*] Sung-Hoon Lee[1*] Tae-Young Lee[1*] Seungjun Moon[2]
Gyeong-Moon Park[1†]

[1]Kyung Hee University, Yongin, Republic of Korea
[2]KLleon Tech., Seoul, Republic of Korea

{jwseo001, sunghoonlee961, slcks1, gmpark}@khu.ac.kr
seungjun.moon@klleon.io

Abstract

*Recent advances in generative models trained on large-scale datasets have made it possible to synthesize high-quality samples across various domains. Moreover, the emergence of strong inversion networks enables not only a reconstruction of real-world images but also the modification of attributes through various editing methods. However, in certain domains related to privacy issues, e.g., human faces, advanced generative models along with strong inversion methods can lead to potential misuses. In this paper, we propose an essential yet under-explored task called generative identity unlearning, which steers the model not to generate an image of specific identity. In the generative identity unlearning, we target the following objectives: (i) preventing the generation of images with a certain identity, and (ii) preserving the overall quality of the generative model. To satisfy these goals, we propose a novel framework, **G**enerative **U**nlearning for Any **IDE**ntity (**GUIDE**), which prevents the reconstruction of a specific identity by unlearning the generator with only a single image. GUIDE consists of two parts: (i) finding a target point for optimization that un-identifies the source latent code and (ii) novel loss functions that facilitate the unlearning procedure while less affecting the learned distribution. Our extensive experiments demonstrate that our proposed method achieves state-of-the-art performance in the generative machine unlearning task. The code is available at https://github.com/KHU-AGI/GUIDE.*

1. Introduction

Recently, 2D or 3D Generative Adversarial Networks (GANs) [4, 19–21] pre-trained on large datasets, *e.g.*, FFHQ [19] or AFHQ [5], have drawn substantial attention due to their remarkable generation performance and highly disentangled representation space. However, their advancements have raised privacy concerns [15], especially regarding the potential misuse of generative models to represent and exploit individual identities. For instance, deepfakes [48, 49] can create very believable images or videos of people in made-up situations, causing major concerns about ethics and privacy.

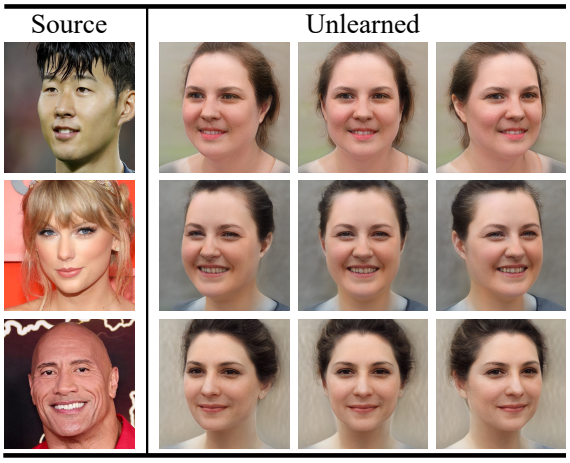

Figure 1. Given a single source image containing a specific identity, we remove that identity from the pre-trained 3D generative adversarial network (e.g., EG3D [4]). Our method effectively unlearns identity even from in-the-wild images where the source image is absent in the pre-training dataset.

To alleviate privacy issues in generative models, machine unlearning task has been actively studied. Machine unlearning involves the process of selectively removing specific knowledge or erasing the influence of certain data from the training dataset of pre-trained models. It is beneficial especially when the data are harmful, private, or biased

*Equal contribution
†Corresponding author

[11, 12, 26, 31]. Despite a focus on discriminative tasks in most machine unlearning research, a few studies have ventured into generative models, attempting to erase high-level concepts such as socially inappropriate content or artistic styles that present copyright challenges [9, 25].

Nevertheless, generative models still exhibit ongoing privacy issues. Even if an identity of someone is not used in the pre-training of the generative models, it can be easily reconstructed in the pre-trained models via GAN inversion models [29, 34, 35, 45, 46, 52]. Furthermore, the reconstructed image can be manipulated or edited easily via image editing methods [30, 39, 40]. To prevent potential exploits of an identity, it is necessary to erase a certain identity from the pre-trained generative models.

To consider the above issue, we introduce an essential task of unlearning any identity from the pre-trained 2D or 3D GANs [4, 20], called *generative identity unlearning*. Unlike typical machine unlearning tasks, which focus on unlearning the training samples our generative identity unlearning task unlearns any identity on pre-trained GANs, even if it was not shown during the pre-training. Our goal is to remove the whole identity associated with a given single image from the generator while minimally impacting the overall performance of the pre-trained model.

To achieve our goal, we propose a novel generative unlearning framework, **G**enerative **U**nlearning for Any **IDE**ntity, named **GUIDE**. GUIDE replaces the source identity with an anonymous target identity, erasing the original identity effectively. To this end, we propose a new exploration method to determine an effective target latent code, called Un-Identifying Face On latent space (UFO). UFO utilizes the GAN inversion method [52] to embed the given identity into the source latent, and then decides the target latent using both the source and the average latent codes. We empirically find that the proposed UFO can identify the promising target to erase any given source identity robustly.

Given the source and target latent code, we update the generator to shift from the source identity to the target identity. To this end, we propose three novel loss functions: (i) *local unlearning loss*, (ii) *adjacency-aware unlearning loss*, and (iii) *global preservation loss*. (i) guides our model directly shifting the source identity to the target identity. (ii) utilizes other latent codes adjacent to the source and target latent codes to effectively unlearn the entire identity from a single image. To minimize side effects from the unlearning process, (iii) additionally regularizes the generator to retain generation performance for latent codes relatively far from the source and target latent codes. Through comprehensive experiments on diverse identities, including *Random*, *InD*, and *OOD*, we confirm that GUIDE can successfully remove the identity of the source image from the pre-trained generative model, and shows qualitatively and quantitatively superior performances.

Our contributions can be summarized as follows:

- For the first time, we propose a novel task, generative identity unlearning, which tackles machine unlearning in generative models in the aspect of privacy protection. In our task, we aim to prevent the pre-trained generative models from synthesizing the given identity by utilizing only a single image.
- For the effective and robust elimination of the identity, we propose a novel method - Un-Identifying Face On Latent Space (UFO). We configure the unlearning procedure by formulating how to represent and shift the identity in the latent space. We find that setting the extrapolated latent code between the source and average latent codes as an optimization target facilitates the unlearning procedure.
- We propose three loss functions - local unlearning loss, adjacency-aware unlearning loss, and global preservation loss to effectively unlearn the identity from the pre-trained model while less affecting the generation performance on other identity.
- We show that our proposed framework, GUIDE achieves state-of-the-art performance both qualitatively and quantitatively, through extensive experiments. We demonstrate that GUIDE can remove the specific identity successfully in the generative models while minimizing the negative effect on other identities.

2. Related Work

Generative Models and Privacy Issue. In image synthesis field, GAN-based generative models have achieved remarkable performance not only in 2D [18–21, 32, 33, 36, 37] but also in 3D [4, 42, 47, 55, 56] domain. The application of various image editing methods [30, 40, 50] to strong generative models, people can easily generate edited images of specific individuals and various artistic styles [38, 41], as well as extract copyrighted content [3] without the permission of the individual or the original creator.

Recently, with the rise of the importance of AI ethics, several works have addressed this issue [9, 25, 28, 44, 53]. ESD [9] erases specific visual concepts from diffusion model by using negative guidance about the undesired concepts. Kumari et al. [25] modifies the conditional distribution of the model a specific target concepts to match the anchor concept. Forget-Me-Not [53] fine-tunes U-Net to minimize each of the intermediate attention associated with the target concepts to remove. Additional works in GANs [28, 44] focus on unlearning specific features, *e.g.*, "Bang", "Hat" or "Beard" rather than forgetting specific identity. The above methods primarily concentrate on the elimination of specific concepts or high-level features. In other words, these cannot preclude models from generating specific individuals while maintaining the generation performance of realistic human faces. Unlike the existing works, our work targets to unlearn only specific individuals without

shifting the overall distribution of generated images.

Machine Unlearning. Machine unlearning aims to selectively forget specific acquired knowledge or diminish the impact of certain training data subsets on a trained model. Since previous research [8, 22, 51] shows that machine learning models might accidentally share private information when faced with certain attacks or inputs, machine unlearning becomes crucial.

While previous machine unlearning is mainly focused on supervised learning tasks [2, 6, 10–14, 43], the interest in unlearning techniques within unsupervised learning, *i.e.*, generative models, is growing [24, 28, 44]. However, most of the existing methods need full dataset access for retraining, which is hard to acquire and computationally expensive [2, 10, 14, 43]. For example, Kong and Chaudhuri [24] utilizes data redaction and augmentation algorithms, which requires a full training dataset. Despite the existence of a feature unlearning model [28] which does not need full dataset access, unlearning only an individual feature is not enough to forget a whole specific identity. To this end, we propose an algorithm that enables forgetting the specific identity only with a single image. Furthermore, our approach distinguishes itself from existing research by applying unlearning to unseen images, enabling the erasure of specific identities without prior exposure to those images.

3. Method

Firstly, in Section 3.1, as shown in Figure 2, we introduce the problem we aim to address, named generative identity unlearning. In Section 3.2, we introduce un-identifying face on latent space, which designates an appropriate target latent for unlearning. Then, in Section 3.3, we introduce latent target unlearning, along with our proposed novel losses to unlearn the generator. The total overview of our method is illustrated in Figure 3.

3.1. Problem Formulation

Given a set of images $\mathbf{x}$ depicting a specific identity, we randomly select a single source image $x_u \in \mathbf{x}$ as an exemplar of the identity. Initially, using off-the-shelf inversion network [52] E corresponding to the unconditional generator, *i.e.*, EG3D [4], we embed x_u to the source latent code w_u in the latent space of EG3D:

$$w_u = E(x_u). \tag{1}$$

Since EG3D [4] consists of the mapping network $Map(\cdot)$, StyleGAN2 [19] backbone $G(\cdot)$ and the neural renderer with a super-resolution module $R(\cdot)$, we can denote the reconstructed image $\hat{x}$ from w_u as following:

$$\hat{x} = R(G(w_u); c), \tag{2}$$

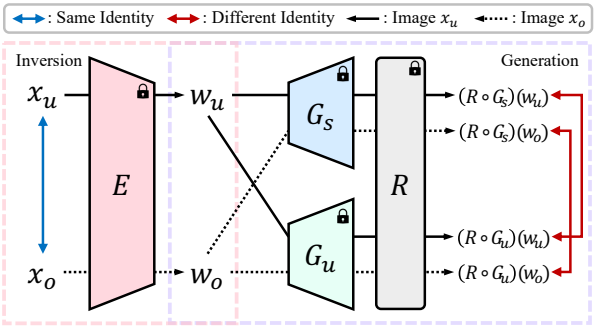

Figure 2. An illustration of *generative identity unlearning*. Upon GUIDE, the identity of the image generated from w_u, *i.e.*, inversion of the source image x_u by inversion network E, should exhibit a distinct identity when passed through the pre-trained generator G_s compared to the unlearned generator G_u. Furthermore, other images x_o, not used in unlearning but sharing the same identity with x_u, also should vary an identity through GUIDE.

where c denotes camera poses. For convenience, we omit the explicit notation of camera poses in this paper, *i.e.*, $\hat{x} = (R \circ G)(w_u)$. We target to derive an unlearned G, *i.e.*, G_u, from the pre-trained EG3D generator G, *i.e.*, G_s, while fixing Map and R. With proper unlearning, an image generated by unlearned EG3D using w_u, *i.e.*, $\hat{x}_u = (R \circ G_u)(w_u)$ should have a distinct identity from the image generated by original EG3D using w_u, *i.e.*, $(R \circ G_s)(w_u)$.

In our task formulation, two considerations are paramount. First, we aim to eliminate the entire identity from the generator only utilizing a single image. To validate this, we incorporate other multiple images $x_o \in \mathbf{x}$ for testing and its corresponding latent code $w_o = E(x_o)$. By utilizing x_o, we can verify whether G_u has successfully unlearned the identity as a whole, rather than just unlearning the specific image x_u. Second, we strive to maintain the generation performance of the pre-trained model. To assess this, we sample multiple images from fixed latent codes using both the unlearned and pre-trained generators. We then estimate the distribution shift between the images generated before and after the unlearning process.

3.2. Un-Identifying Face On Latent Space

The successfully unlearned model should not generate the image with the identity of $\mathbf{x}$, even when w_u is used as a latent. Consequently, we initiate our approach to manipulate $\hat{x}_u$ to be another image rather than the image with identity in $\mathbf{x}$ by unlearning G. To design the objective function for unlearning, we first need to establish the objective for unlearning, which involves defining the target image $\hat{x}_t$, derived from w_t, which the unlearned image $\hat{x}_u$, derived from w_u, should mimic after unlearning. While there exist various options for setting $\hat{x}_t$, *e.g.*, a random face or even a non-human image, we choose the mean face generated by

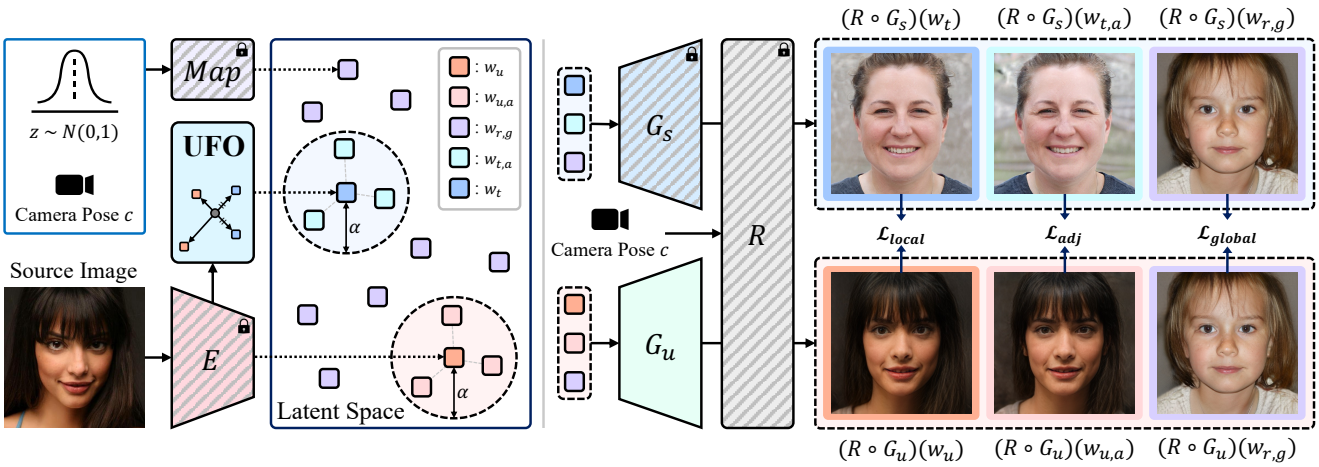

Figure 3. An overview of GUIDE. Starting with a source image, we employ a GAN inversion network E, specifically GOAE [52], to embed this image into the latent space of a pre-trained generative model, namely EG3D [4], obtaining the source latent code w_u. The target latent code w_t is designated through the UFO process. To facilitate identity removal in w_u, we shift its identity to match that of w_t with our Latent Target Unlearning (LTU) process. Three loss functions of LTU are designed for this purpose: (i) The generator is optimized to produce an image from the source latent code, denoted as $(R \circ G_u)(w_u)$, that is similar to the image from the target latent code, represented as $(R \circ G_s)(w_t)$. (ii) To achieve unlearning across the entire identity, we consider latent codes near both the source and target latent codes, denoted as $w_{u,a}$ and $w_{t,a}$, respectively. (iii) To prevent model corruption during the unlearning process, we additionally sample latent codes from a random noise vector, represented as $w_{r,g}$, and optimize G_u to preserve its generation ability on $w_{r,g}$.

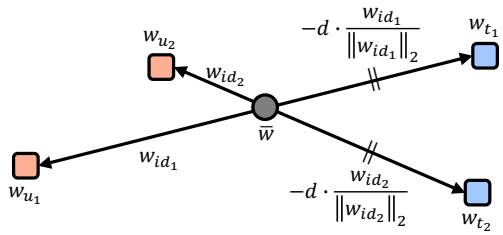

Figure 4. An illustration of Un-identifying Face On Latent Space (UFO). We define the identity of the source latent code by subtract it from the average latent code. We set the target latent code for our unlearning process by measuring an extrapolation between the source and average latent code with a fixed distance d.

the mean latent $\bar{w}$ of Map, *i.e.*, $(R \circ G_s)(\bar{w})$, as $\hat{x}_t$. This setting is inspired by inversion methods [1, 29] which apply the identity to the mean face through inversion stages. Since our goal is to de-identify the image, we argue that it is intuitive and reasonable to send the image back to the mean face, which is opposite to the inversion.

However, this simple approach might be problematic when w_u is close to $\bar{w}$, where the unlearned image might still resemble the original identity. To this end, we propose a novel method named Un-identifying Face On latent space (UFO), which can set $\hat{x}_t$ robustly regardless of the distance between w_u and $\bar{w}$, as shown in Figure 4. The following are the processes of UFO: First, in the *de*-identification process, we extract the identity latent $w_{id} = w_u - \bar{w}$ by subtracting $\bar{w}$ from w_u. Then we propose a process of *en*-identification, setting the target in the opposite direction of the existing w_{id} to foster a more pronounced change in identity, creating an entirely new identity. In other words, we chose $\bar{w}$ as the stopover point and established the final target point at the extrapolation of w_u and $\bar{w}$. The finalized target point can be expressed as:

$$w_t = \bar{w} - d \cdot \frac{w_{id}}{\|w_{id}\|_2}, \tag{3}$$

where d is defined as the distance that balances un-identification with preservation of the source distribution, as determined by our empirical results. Finally, we can set $\hat{x}_t = (R \circ G_s)(w_t)$. We empirically demonstrate that whether the given identity is in close proximity to the average latent code within the latent space or not, UFO sets the desirable target latent code for the generative unlearning.

3.3. Latent Target Unlearning

After setting $\hat{x}_t$, we need to robustly design the unlearning process that can effectively make $\hat{x}_u$ to be similar to $\hat{x}_t$, while maintaining the generation performance of G_u. We coin this process as Latent Target Unlearning (LTU), which targets to unlearn images from the specific latent while keeping generation results from other latent codes. LTU utilizes the following three losses to achieve this goal:

Local Unlearning Loss. To force $\hat{x}_u$ to be $\hat{x}_t$, we use the widely-used reconstruction losses, *i.e.*, Euclidean loss $\mathcal{L}_2$,

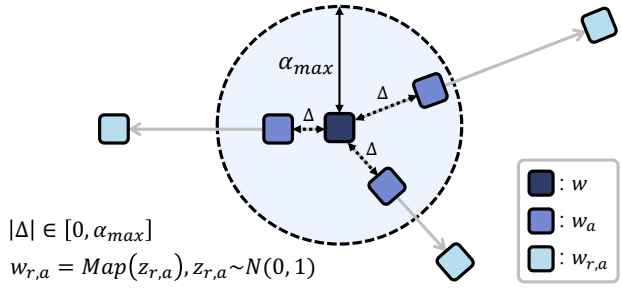

Figure 5. An illustration of determining latent codes near a latent code w in adjacency-aware unlearning loss. We first sample a latent code $w_{r,a}$ which is derived from a random noise vector $z_{r,a}$ via the mapping network $Map(\cdot)$, *i.e.* $w_{r,a} = Map(z_{r,a})$. Next, we compute the direction between w and $w_{r,a}$, and we scale it to fall within range between 0 and α_{max}. This yields the distance vector Δ to compute the adjacent latent code $w_a = w + \Delta$.

perceptual loss $\mathcal{L}_{per}$ [54], and identity loss $\mathcal{L}_{id}$ [7]. Using $\mathcal{L}_{recon}$, we compare the tri-plane features $F_u = G_u(w_u)$ and $F_t = G_s(w_t)$, derived from source and target latent codes, respectively. The local unlearning loss is defined as:

$$\begin{aligned}\mathcal{L}_{local}(\hat{x}_u, \hat{x}_t) = \lambda_{L2}\mathcal{L}_{L2}(F_u, F_t) \\ + \lambda_{per}\mathcal{L}_{per}(\hat{x}_u, \hat{x}_t) + \lambda_{id}\mathcal{L}_{id}(\hat{x}_u, \hat{x}_t).\end{aligned} \tag{4}$$

By adopting $\mathcal{L}_{local}$, we can successfully un-identify the given source identity in $\hat{x}_t$.

Adjacency-Aware Unlearning Loss. The above equation considers only one pair of source and target latent codes. However, images of a similar identity to the source identity can be obtained by introducing marginal perturbations to the latent code. For the successful unlearning of the given identity, we need to consider the neighborhood of both the source and the target latent codes. Consequently, as shown in Figure 5, we sample N_a latent codes in the vicinity of the w_u. Specifically, with the scale α^i sampled from the uniform distribution with hyperparameter α_{max}, *i.e.* $\alpha^i \sim \mathcal{U}(0, \alpha_{max})$, we define the distances Δ to compute the adjacent latent codes as:

$$\Delta = \{\alpha^i \cdot \frac{w^i_{r,a} - w_u}{\|w^i_{r,a} - w_u\|_2}\}^{N_a}_{i=1}, \tag{5}$$

where $w^i_{r,a}$ is a latent code sampled from the random noise vector $z^i_{r,a}$. Using these distances Δ, we can compute N_a latent codes for both the source and the target latent codes. Similar to the local unlearning loss, we optimize the generated tri-plane features and images from $w^i_{u,a} = w_u + \Delta^i$ and $w^i_{t,a} = w_t + \Delta^i$ to be similar:

$$\hat{x}^i_{u,a} = R(F^i_{u,a}), \hat{x}^i_{t,a} = R(F^i_{u,a}), \tag{6}$$

$$\mathcal{L}_{adj}(w_u, w_t) = \frac{1}{N_a}\sum_{i=1}^{N_a}\mathcal{L}_{local}(\hat{x}^i_{u,a}, \hat{x}^i_{t,a}), \tag{7}$$

where $F^i_{u,a} = G_u(w^i_{u,a}), F^i_{t,a} = G_s(w^i_{t,a})$ denotes for tri-plane features, and $\mathcal{L}_{local}$ in Equation 4. From $\mathcal{L}_{adj}$, we can further consider possible variations of the source identity.

Global Preservation Loss. While the local unlearning loss and adjacency-aware unlearning loss mentioned above facilitate the removal of the source identity, we propose a global preservation loss to mitigate side effects arising from these unlearning loss functions. In the global preservation loss, we constrain the generator to maintain generation performance for latent codes that are relatively distant from both the source and target latent codes.

To be precise, we sample N_g latent codes $\{w^i_{r,g}\}^{N_g}_{i=1}$ from random noise vectors $\{z^i_{r,g}\}^{N_g}_{i=1}$. We ensure that these do not overlap with the adjacent latent codes used in the adjacency-aware unlearning loss. Unlike the unlearning loss functions, we find that adopting only L_{per} achieves a balanced performance between identity shift and model preservation. The global preservation loss is computed as:

$$\begin{aligned}\hat{x}^i_{u,g} &= (R \circ G_u)(w^i_{r,g}), \\ \hat{x}^i_{s,g} &= (R \circ G_s)(w^i_{r,g}), \\ \mathcal{L}_{global}(G_u, G_s) &= \frac{1}{N_g}\sum_{i=1}^{N_g}\mathcal{L}_{per}(\hat{x}^i_{u,g}, \hat{x}^i_{s,g}).\end{aligned} \tag{8}$$

In summary, our final objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{local} + \lambda_{adj}\mathcal{L}_{adj} + \lambda_{global}\mathcal{L}_{global}. \tag{9}$$

4. Experiments

4.1. Experimental Setup

Baseline. Since we propose generative identity unlearning task for the first time, to evaluate the effectiveness of GUIDE, we constructed a simple baseline. In the baseline, we used the target latent code as the average one for the unlearning. During the unlearning, we updated the pre-trained generator using $\mathcal{L}_{local}$ as described in Equation 4.

Implementation Details. We built GUIDE based on the 3D generative adversarial network [4] pre-trained on FFHQ dataset [19]. We used GOAE [52] as a GAN inversion network to obtain the latent code from an image. The image resolution we used in our experiments is 512x512 with a rendering resolution of 128x128. We used Adam optimizer

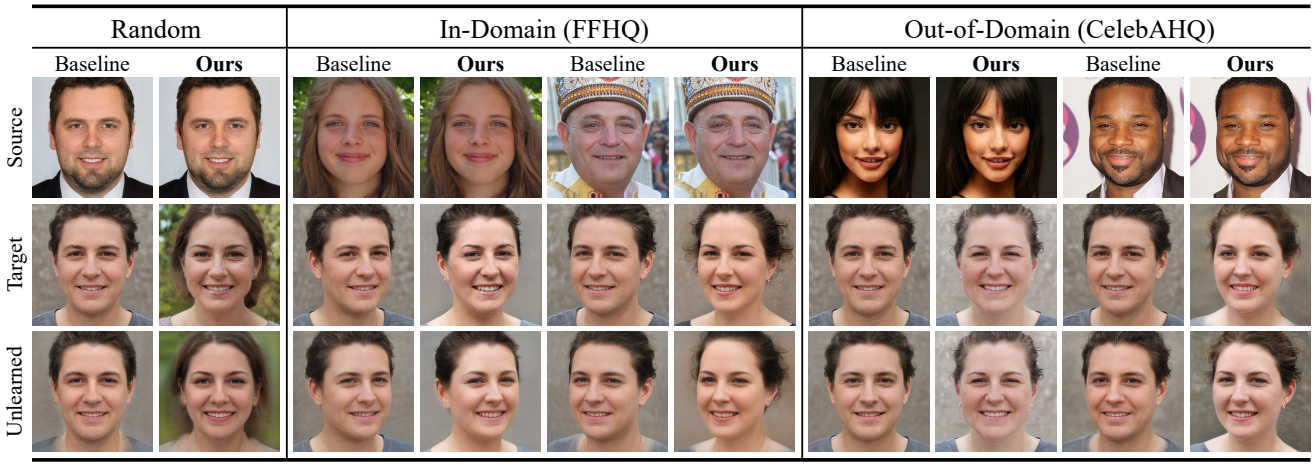

Figure 6. Qualitative results of GUIDE and the baseline in generative identity unlearning task. For the given source image each (the first row), GUIDE and the baseline tried to erase the identity in the pre-trained generator. The images in the second and third row are the target and unlearned images, respectively.

[23] with a learning rate of 10^{-4} in the unlearning procedure. The hyperparameters used in the experiments were: $d = 30$, $\alpha_{max} = 15$, $\lambda_{L2} = 10^{-2}$, $\lambda_{per} = 1$, $\lambda_{id} = 10^{-1}$, $N_a = N_g = 2$, and $\lambda_{adj} = \lambda_{global} = 1$.

Dataset and Scenarios. We evaluated GUIDE in three scenarios: *Random*, where we set an unlearning target image from a randomly sampled noise vector; *InD* (*in-domain*), where we sampled an image from the FFHQ dataset [19] used for pre-training; and *OOD* (*out-of-domain*), where the unlearning target image was sampled from the CelebAHQ dataset [18]. For *InD* and *OOD*, we used the GAN inversion network to obtain corresponding latent codes. For *OOD* scenario, we also conducted *multi-image* test since there were multiple images with a same identity in CelebAHQ. On the other hand, we performed only *single-image* test in the *Random* and *InD* scenarios.

Evaluation Metrics. We evaluated GUIDE on two key aspects. Firstly, we estimated the efficacy of our approach in preventing the generator from producing images similar to the unlearning target. We quantitatively measured similarity of identities (ID) using face recognition network, CurricularFace [17], between images generated from the same latent codes before and after unlearning. Moreover, we utilized $\text{ID}_{\text{others}}$ to estimate the erasure of a identity from images which are unseen during training but containing the same identity of the source image. Secondly, we assessed whether our method preserves overall generation performance using the Fréchet Inception distance (FID) score [16]. Different from the existing usages, we utilized two variants of FID. First, we evaluated the distribution shift of generated images the pre-trained generator and the unlearned generator via FID_{pre}. Furthermore, we measured the distribution shift with respect to the real FFHQ images, which we denoted as $\Delta\text{FID}_{\text{real}}$.

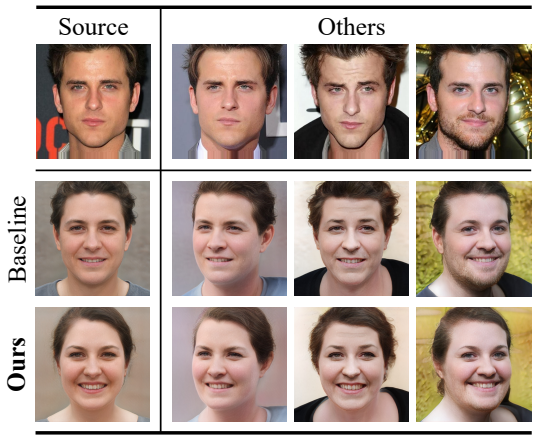

Figure 7. Qualitative results of GUIDE and the baseline on a multi-image test using CelebAHQ dataset. We additionally utilized images that are unseen during unlearning, to show how thoroughly erase the given identity.

4.2. Main Experiment

4.2.1 Qualitative Results

We conducted a comparative analysis of GUIDE against the baseline in the generative identity unlearning task. Initiating from the provided source image, we aimed to eliminate the identity within the pre-trained generator, as illustrated in Figure 6. We presented the resulting unlearned image, along with the target image optimized in our loss functions. Notably, GUIDE effectively erases identities whether synthetic, presented during pre-training, or unseen.

Methods	Random			In-Domain (FFHQ)			Out-of-Domain (CelebAHQ)		
	ID (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)	ID (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)	ID (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
Baseline	0.19 ± 0.09	11.73 ± 2.74	7.46 ± 2.20	0.16 ± 0.07	9.00 ± 1.15	4.15 ± 1.18	0.12 ± 0.06	9.52 ± 1.53	4.75 ± 0.89
+ extrapolated w_t	**0.12 ± 0.06**	14.28 ± 3.34	9.63 ± 2.53	0.05 ± 0.06	12.78 ± 1.82	6.76 ± 1.41	0.02 ± 0.05	13.02 ± 3.20	7.31 ± 1.98
+ $\mathcal{L}_{adj}$	0.14 ± 0.07	19.65 ± 4.90	13.94 ± 3.59	**0.04 ± 0.06**	13.53 ± 2.08	7.35 ± 1.70	**0.01 ± 0.05**	13.63 ± 3.52	7.83 ± 2.19
+ $\mathcal{L}_{global}$ (**GUIDE**)	0.14 ± 0.06	**10.80 ± 2.70**	**6.64 ± 1.60**	0.06 ± 0.06	**8.00 ± 1.20**	**3.05 ± 0.81**	0.03 ± 0.05	**7.88 ± 1.96**	**3.34 ± 1.10**

Table 1. Quantitative results of GUIDE and the baseline in the generative identity unlearning task, tested in a single-image setting using one image per identity. Starting from the baseline, we gradually introduced components of GUIDE.

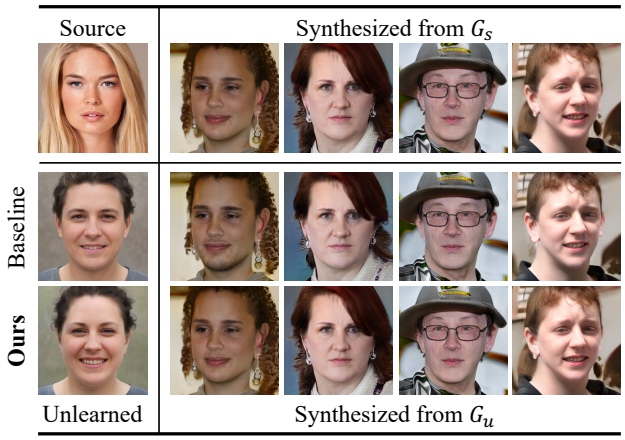

Figure 8. Qualitative comparison between GUIDE and the baseline on the preservation of the generation quality of other identities. GUIDE generates images almost identical to those synthesized by G_s, whereas the baseline often results in noticeable changes, *e.g.*, beard shape, hairstyle change, hat.

To evaluate the thoroughness of identity removal, we performed a multi-image test using identities from the CelebAHQ dataset. This test involved assessing the ID similarity not only for the unlearned image derived from the source image but also for other images sharing the same identity. As shown in Figure 7, GUIDE showed superior generalization for unseen images compared to the baseline. This improvement is attributed to the adjacency-aware unlearning, which facilitated the unlearning process not just for the given images but also for their neighborhood.

In Figure 8, we conducted an experiment to assess the effect of the unlearning process on other identities. While the baseline had a significant impact on other identities through the unlearning, GUIDE showed a relatively lesser effect. We attribute this to the global preservation loss, which constrained the distribution shift on other latent codes.

4.2.2 Quantitative Results

In Table 1, we compared GUIDE to the baseline by gradually applying the components of GUIDE. By configuring w_t through extrapolation, we achieved performance improvements in ID similarity across three scenarios. Notably, we observed a significant difference in ID similarities in the random scenario, indicating that in cases where a latent code was close to $\bar{w}$, *i.e.*, as in the random scenario, there was insufficient removal of identity. The effectiveness of employing an extrapolation between the source latent code and the average latent code was evident in such instances.

Methods	ID (↓)	$\text{ID}_{\text{others}}$ (↓)	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
Baseline	0.12 ± 0.06	0.28 ± 0.08	9.52 ± 1.53	4.75 ± 0.89
+ extrapolated w_t	0.02 ± 0.05	0.15 ± 0.07	13.02 ± 3.20	7.31 ± 1.98
+ $\mathcal{L}_{adj}$	**0.01 ± 0.05**	**0.14 ± 0.07**	13.63 ± 3.52	7.83 ± 2.19
+ $\mathcal{L}_{global}$ (**GUIDE**)	0.03 ± 0.05	0.17 ± 0.08	**7.88 ± 1.96**	**3.34 ± 1.10**

Table 2. Quantitative results of GUIDE and the baseline in the generative identity unlearning in a multi-image setting, *i.e.*, using a single image for unlearning and the other images for testing. We used CelebAHQ dataset for this test.

The adjacency-aware unlearning loss further enhanced the unlearning an identity. This loss was designed to cover the vicinity of the source latent code, thereby promoting unlearning on the source latent code itself. Finally, the application of the global preservation loss effectively reduced the estimated distribution shift using FID_{pre} and $\Delta\text{FID}_{\text{real}}$.

Moreover, we conducted a multi-image test in an *OOD* scenario. In this particular experiment, we introduced additional metric - $\text{ID}_{\text{others}}$ aimed at quantifying ID similarities for the unseen images associated with the source identity. As presented in Table 2, the introduction of the adjacency-aware unlearning loss resulted in a remarkable improvement in $\text{ID}_{\text{others}}$, emphasizing the effectiveness of this unlearning approach for handling unseen images.

4.3. Ablation Study

Effect of d in Determination of w_t. We conducted an ablation study by comparing target images derived from varied values of d. Setting d to 0 denotes utilizing the $\bar{w}$ as w_t in the unlearning process. For $d < 0$, we designated w_t as an interpolated latent code between the w_s and the $\bar{w}$. Conversely, for $d > 0$, we employed an extrapolated w_t, as detailed in Section 3.2. As illustrated in Figure 9, when $d < 0$, the target image closely aligns with the given source images. However, as d deviates from 0, the quality of the target image rapidly deteriorates, resulting in a pronounced collapse in the distribution of the pre-trained generator. Consequently, the effectiveness of unlearning with such target images diminishes in removing identity from the

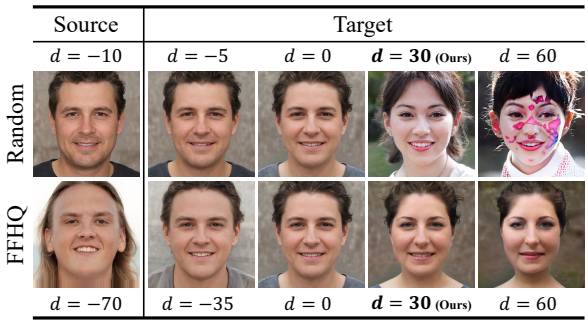

Figure 9. Ablation study to figure out the effectiveness of d. We visualized target images corresponding to each source image with different values of d. The target images were generated using target latent codes derived from interpolated latent codes, the average latent code ($d = 0$), or extrapolated latent codes ($d > 0$). Interpolation and extrapolation were carried out between the source and the average latent code. In the case of interpolation, the center between the source and the average latent code was computed.

α_{max}	ID (↓)	$\text{ID}_{\text{others}}$ (↓)
0	0.1205 ± 0.0603	0.2754 ± 0.0791
10	0.0892 ± 0.0620	0.2123 ± 0.0762
15	**0.0878** ± **0.0375**	**0.2094** ± **0.0692**
20	0.0900 ± 0.0538	0.2105 ± 0.0924
30	0.0926 ± 0.0561	0.2111 ± 0.0653

Table 3. Ablation study to figure out the effectiveness of $\mathcal{L}_{adj}$ and α_{max}. We compared the performance based on how successfully the given identity was erased, using ID and $\text{ID}_{\text{others}}$ metric. The row where $\alpha_{max} = 0$ denotes the baseline. We used CelebAHQ dataset in this experiment.

source images. Setting d to 0 might suggest the use of $\bar{w}$ as an effective target for erasing identity. However, our ablation studies indicate that when the source image closely aligns with $\bar{w}$, the unlearning procedure fails to thoroughly eliminate the identity. Conversely, when $d > 0$, w_t contains a distinct identity compared to the source image while maintaining a consistent distance from $\bar{w}$. Among the instances where $d > 0$, our ablation studies reveal that setting $d = 30$ achieves a balanced performance between effective unlearning and preservation of the generation performance of the pre-trained model.

Effect of α_{max} in $\mathcal{L}_{adj}$. In Table 3, we scrutinized the effectiveness of the adjacency-aware unlearning loss. To ensure a fair comparison, we employed $\bar{w}$ as w_t in this experiment, and we used $\mathcal{L}_{local}$ and $\mathcal{L}_{adj}$ in the unlearning procedure. Rows corresponding to $\alpha_{max} = 0$ represent experimental results without the incorporation of $\mathcal{L}_{adj}$ in the unlearning procedure. The introduction of $\mathcal{L}_{adj}$ resulted in consistent performance gains in $\text{ID}_{\text{others}}$. This observation highlights the efficacy of considering not only the pair of source and target latent codes but also their surroundings for unlearning the entire identity.

$\mathcal{L}_{local}$	$\mathcal{L}_{global}$	FID_{pre} (↓)	$\Delta\text{FID}_{\text{real}}$ (↓)
✓		9.52 ± 1.53	4.75 ± 0.89
✓	✓	**4.63** ± **0.43**	**1.48** ± **0.29**

Table 4. Ablation study to figure our the effectiveness of $\mathcal{L}_{global}$. We compared how preserved the performance of the pre-trained model through the unlearning process, via FID_{pre} and $\Delta\text{FID}_{\text{real}}$. We used CelebAHQ dataset in this experiment.

Effect of $\mathcal{L}_{global}$. To assess the effectiveness of the global preservation loss, a similar experiment was conducted as the previous experiment, *i.e.*, setting $\bar{w}$ as w_t. The results are presented in Table 4. The application of $\mathcal{L}_{global}$ demonstrated consistent performance improvements in both FID_{pre} and $\Delta\text{FID}_{\text{real}}$. This suggests that imposing constraints on the generator to maintain its generation performance in latent codes distant from our primary focus is effective in reducing distribution shifts in generative models.

5. Conclusion

In this paper, we introduced a novel task, referred to as generative identity unlearning, designed to address privacy concerns in pre-trained generative adversarial networks. This task requires thoroughly removing the identity of a single source image from the pre-trained generator. To achieve this, we proposed a new framework, GUIDE (Generative Unlearning for any IDEntity). To unlearn the single identity, we first defined the target latent code via extrapolation, moving away from the average latent by the pre-defined distance in the direction from the source to the average latent. Using this, our GUIDE successfully unlearned the given identity via Latent Target Unlearning (LTU), which optimized the pre-trained model to preserve the overall generative ability but not to generate the same identity within the local space. Experimental results demonstrated the effectiveness of GUIDE with promising outcomes. We anticipate that our work will be widely applied in research or the industry field, providing users with a sense of freedom from privacy concerns through identity removal.

Acknowledgement

This work was supported by MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2023-00258649) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and in part by NRF-2023S1A5A2A21083590, and in part by the IITP grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068, and by the IITP grant funded by the Korea government (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)).

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6711–6720, 2021. 4

[2] Thomas Baumhauer, Pascal Schöttle, and Matthias Zeppelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *Machine Learning*, 111(9):3203–3226, 2022. 3

[3] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 2

[4] 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, 3, 4, 5

[5] 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

[6] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 2023. 3

[7] 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. 5

[8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015. 3

[9] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision (ICCV)*, pages 2426–2436, 2023. 2

[10] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[11] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9304–9312, 2020. 2

[12] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–398. Springer, 2020. 2

[13] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 792–801, 2021.

[14] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330, 2021. 3

[15] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017. 1

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020. 6

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability and variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2, 6

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 5, 6

[20] 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. 2

[21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 1, 2

[22] Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi. Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system. *IEEE Transactions on Information Forensics and Security*, 17:357–372, 2022. 3

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 6

[24] Zhifeng Kong and Kamalika Chaudhuri. Data redaction from pre-trained gans. In *Proceedings of the 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 638–677. IEEE, 2023. 3

[25] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision (ICCV)*, pages 22691–22702, 2023. 2

[26] Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N Ravi. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10422–10431, 2022. 2

[27] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11731–11741, 2023.

[28] Saemi Moon, Seunghyuk Cho, and Dongwoo Kim. Feature unlearning for generative models via implicit feedback. *arXiv preprint arXiv:2303.05699*, 2023. 2, 3

[29] Seung-Jun Moon and Gyeong-Moon Park. Interestyle: Encoding an interest region for robust stylegan inversion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 460–476. Springer, 2022. 2, 4

[30] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 2

[31] Alexandra Peste, Dan Alistarh, and Christoph H Lampert. Ssse: Efficiently erasing samples from trained machine learning models. corr. *arXiv preprint arXiv:2107.03860*, 1 (3), 2021. 2

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8821–8831. PMLR, 2021. 2

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[34] 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. 2

[35] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 2

[36] 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 (CVPR)*, pages 10684–10695, 2022. 2

[37] Juwon Seo, Ji-Su Kang, and Gyeong-Moon Park. Lfs-gan: Lifelong few-shot image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11356–11366, 2023. 2

[38] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023. 2

[39] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252, 2020. 2

[40] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. 2

[41] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058, 2023. 2

[42] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20991–21002, 2023. 2

[43] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3

[44] Piyush Tiwary, Atri Guha, Subhodip Panda, et al. Adapt then unlearn: Exploiting parameter space semantics for unlearning in generative adversarial networks. *arXiv preprint arXiv:2309.14054*, 2023. 2, 3

[45] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 2

[46] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11379–11388, 2022. 2

[47] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems*, 35:36188–36201, 2022. 2

[48] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22658–22668, 2023. 1

[49] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22412–22423, 2023. 1

[50] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9036–9045, 2019. 2

[51] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30 (9):2805–2824, 2019. 3

[52] 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. 2, 3, 4, 5

[53] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 2

[54] 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. 5

[55] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Yi Yang, and Tat-Seng Chua. Multi-view consistent generative adversarial networks for compositional 3d-aware image synthesis. *International Journal of Computer Vision*, pages 1–24, 2023. 2

[56] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 18–35. Springer, 2022. 2