Paper Review

Generative Unlearning for Any Identity

YeongHyeon Park

Department of Electrical and Computer Engineering

SungKyunKwan University

Generative Unlearning for Any Identity

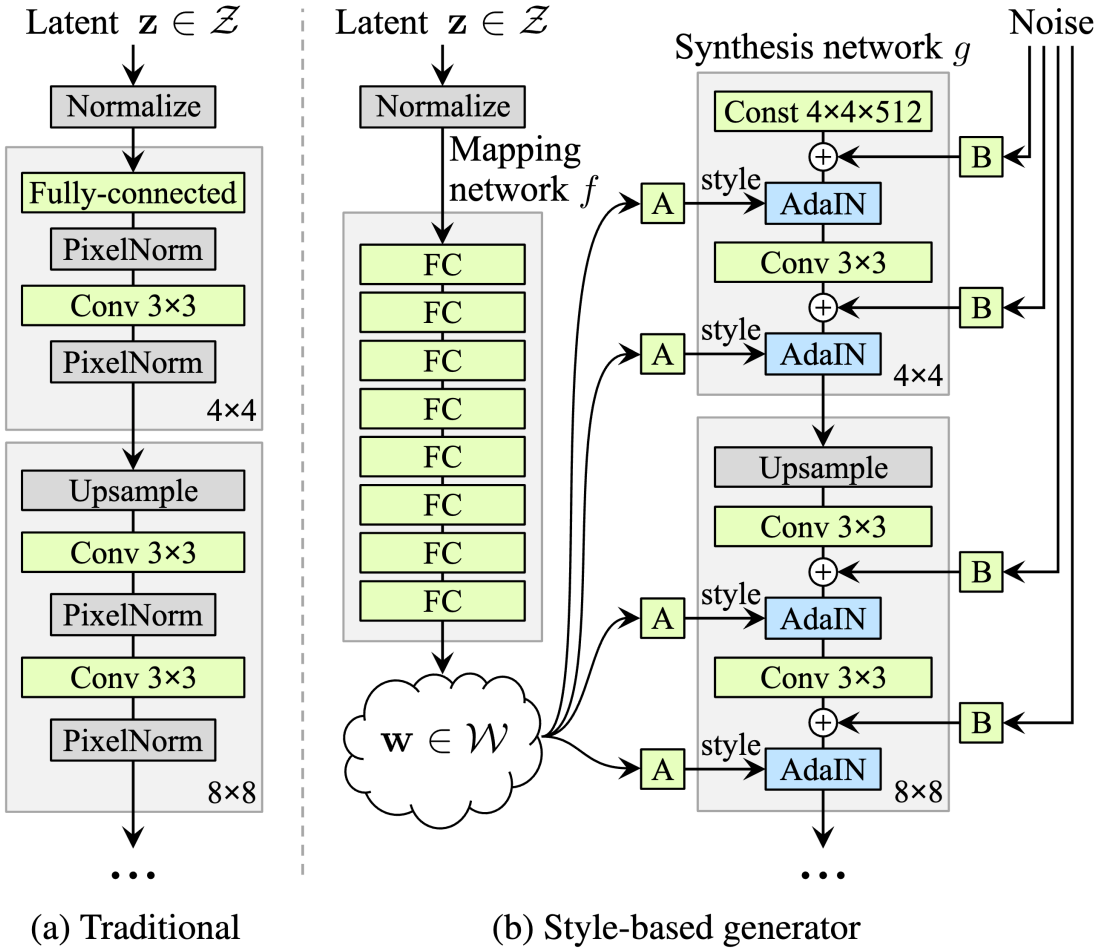
Juwon Seo[1*] Sung-Hoon Lee[1*] Tae-Young Lee[1*] Seungjun Moon[2]
Gyeong-Moon Park[1†]

[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

Warm up

StyleGAN



(a) Traditional (b) Style-based generator

EG3D

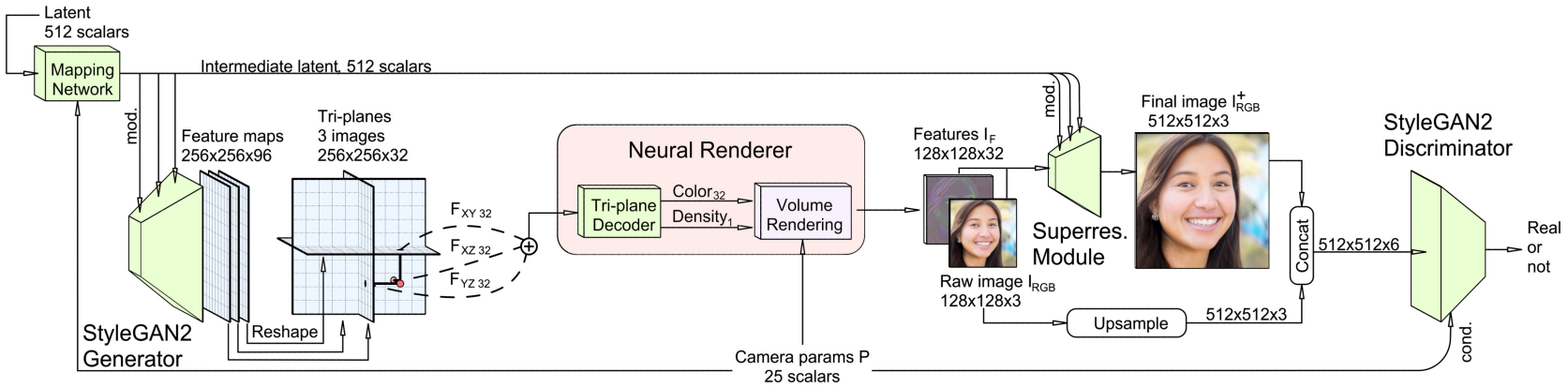


Figure 4. Our 3D GAN framework comprises several parts: a pose-conditioned StyleGAN2-based feature generator and mapping network, a tri-plane 3D representation with a lightweight feature decoder, a neural volume renderer, a super-resolution module, and a pose-conditioned StyleGAN2 discriminator with dual discrimination. This architecture elegantly decouples feature generation and neural rendering, allowing the use of a powerful StyleGAN2 generator for 3D scene generalization. Moreover, the lightweight 3D tri-plane representation is both expressive and efficient in enabling high-quality 3D-aware view synthesis in real-time.

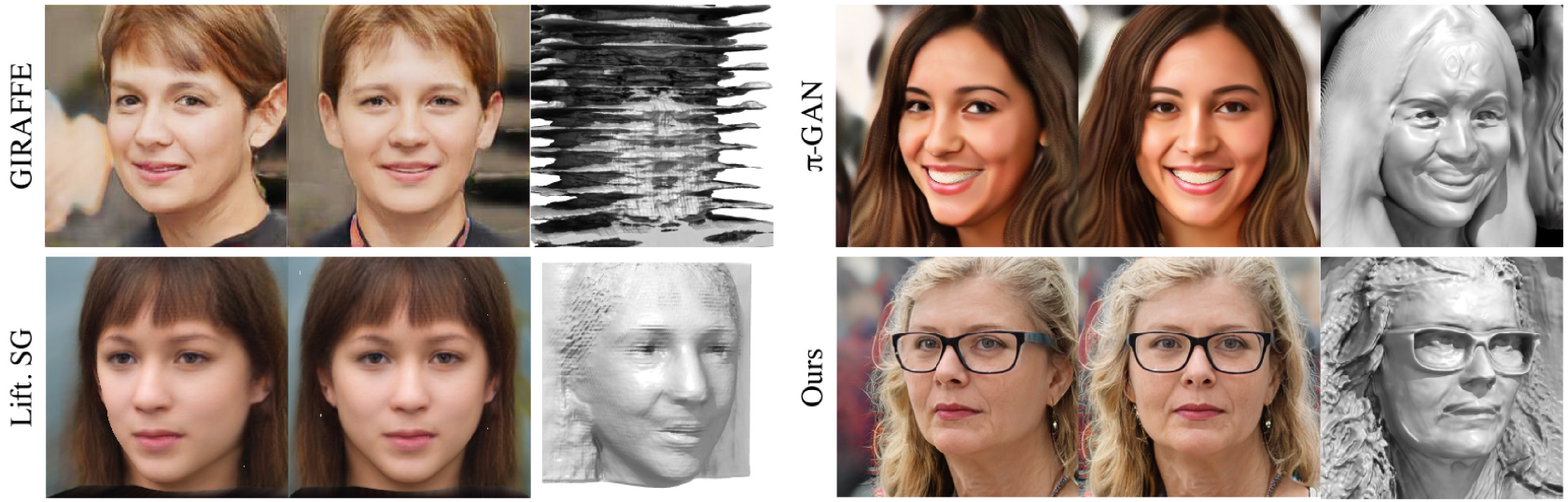


Figure 7. Qualitative comparison between GIRAFFE, π-GAN, Lifting StyleGAN, ours, with FFHQ at 256^2. Shapes are iso-surfaces extracted from the density field using marching cubes. We inspected the underlying 3D representations of GIRAFFE and found that its over-reliance on image-space approximations significantly harms the learning of the 3D geometry.

GOAE

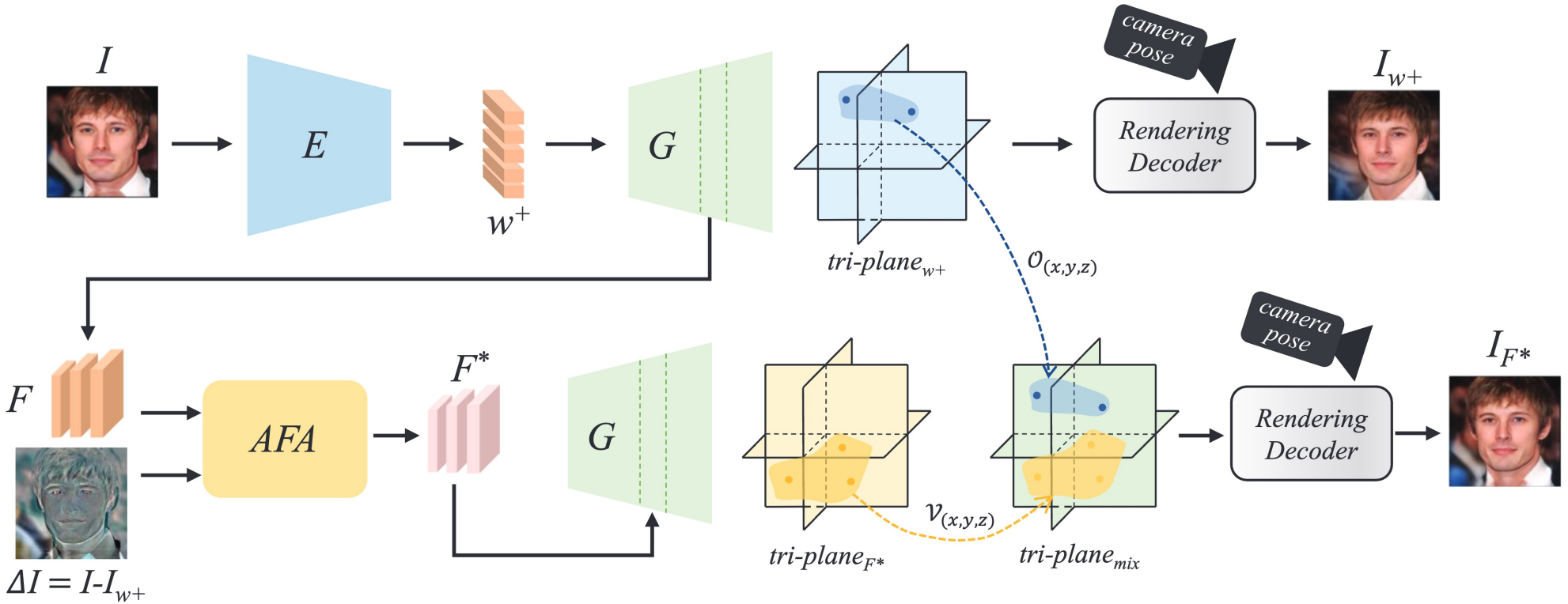


Figure 4: **Overview of our method**. Our framework could be divided into two parts. (1) $\mathcal{W}$ space inversion. We design an encoder E to invert input image I into w^+ latent codes. The w^+ latent codes are fed into a pre-trained EG3D generator G to get $tri\text{-}plane_{w+}$ and rendered into reconstruction image I_{w+}. (2) Complement the $\mathcal{F}$ space. We calculate the image residual ΔI between the input image and its reconstruction and propose AFA module to refine the F latent maps. The modified latent maps F^* are transformed into $tri\text{-}plane_{mix}$ by occlusion-aware mix and rendered into the fine detailed inversion image I_{F^*}.

Summary of GUIDE

Generative **U**nlearning for Any **IDE**ntity

Summaries

Motivation and solution

- **Motivation**: privacy issues of generative models

- **Solution**: unlearning target identity of the pre-trained face generation model

Contributions (authors said)

- Propose a novel task of privacy protection
- UFO method to shift the given identity in the latent space
- Three loss functions of LTU to forget a given identity while preserving high-quality face image generation
- Achieves state-of-the-art while minimizing the negative effect on other identities

Strengths

- Powerful unlearning method with only a single image
  - Completely forgets specific identity through Adjacency-aware unlearning
- Clear problem definition
- Combination of the latest technologies (Not seem to be a completely new concept)
  - UFO: extrapolation method
  - LTU: L_2 loss, perceptual loss, and identity loss

Weaknesses

- Short analysis of experimental results

GUIDE

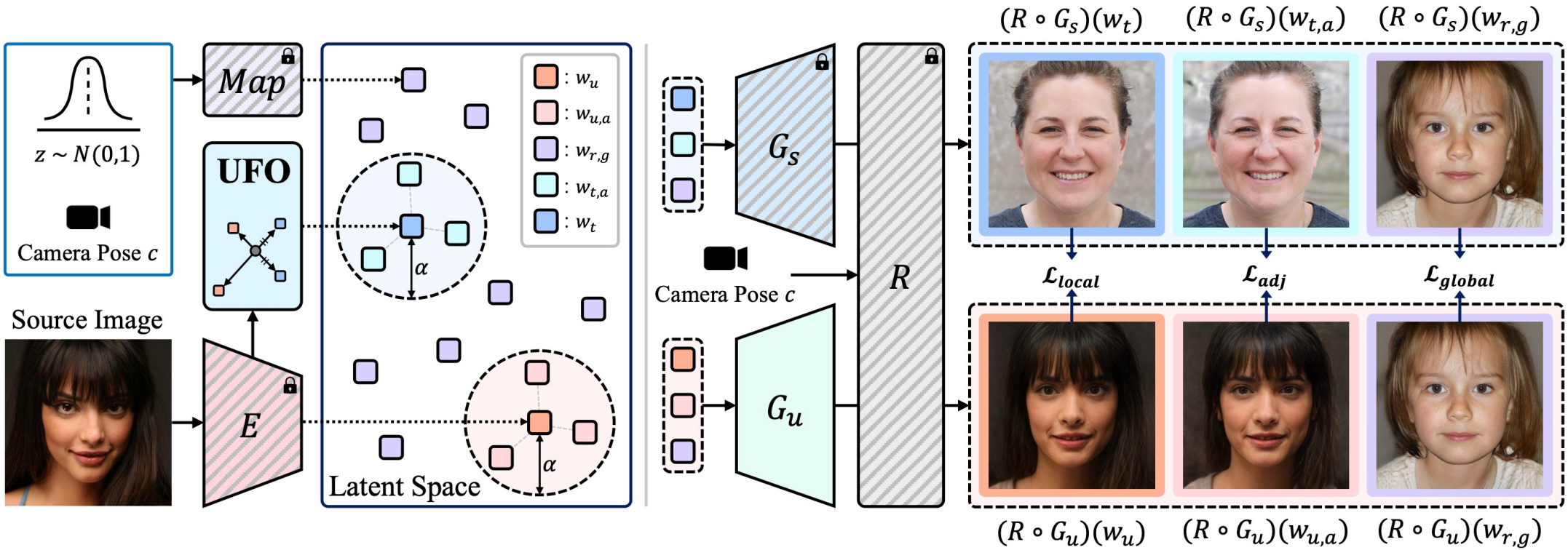
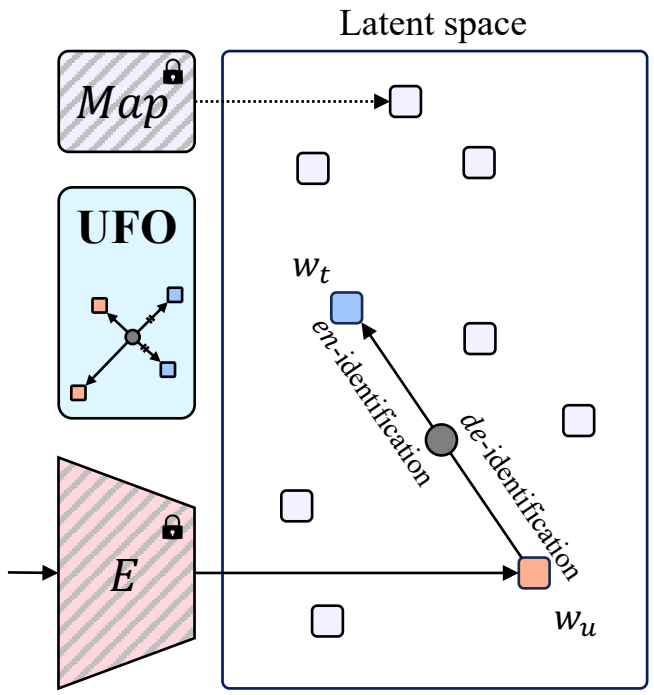
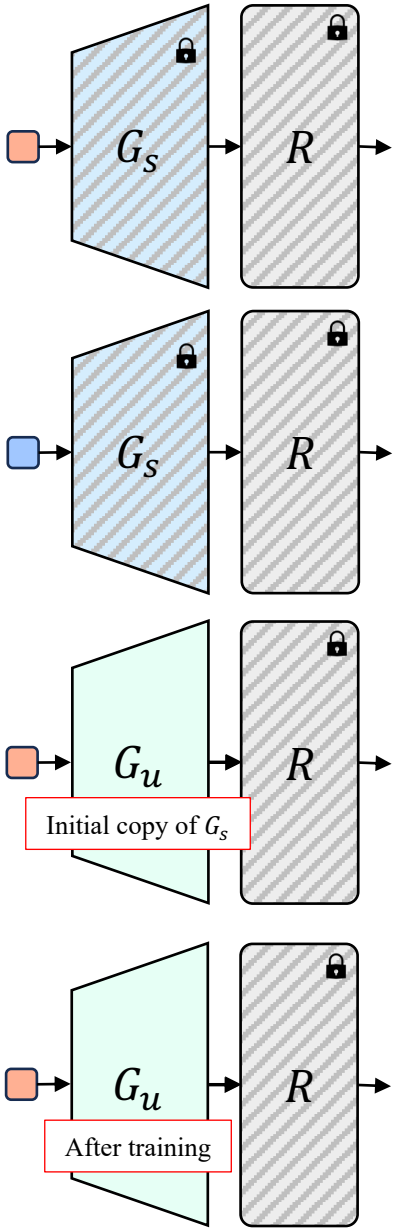


Figure 3. An overview of GUIDE. Starting with a source image, we employ a GAN inversion network E, specifically GOAE [52], to embed this image into the latent space of a pre-trained generative model, namely EG3D [4], obtaining the source latent code w_u. The target latent code w_t is designated through 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 $w_{r,g}$.

GUIDE
Encoding and Generation



w_u: latent of source image (for unlearning)
w_t: target anonymized latent
▢ : known identities



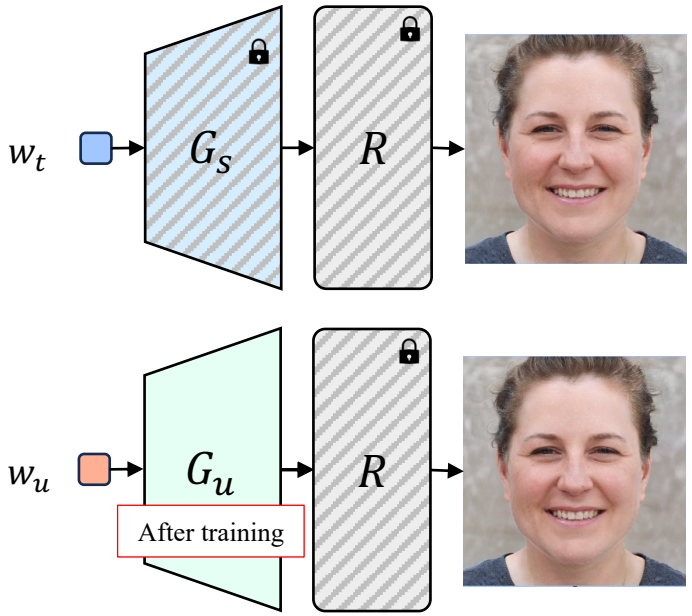
Identity preservation
High-quality generation

Anonymous identity (target for training)
High-quality generation

Identity preservation (before training)
Still high-quality generation

Identity removed (after training)
Preservation of high-quality generation

Question

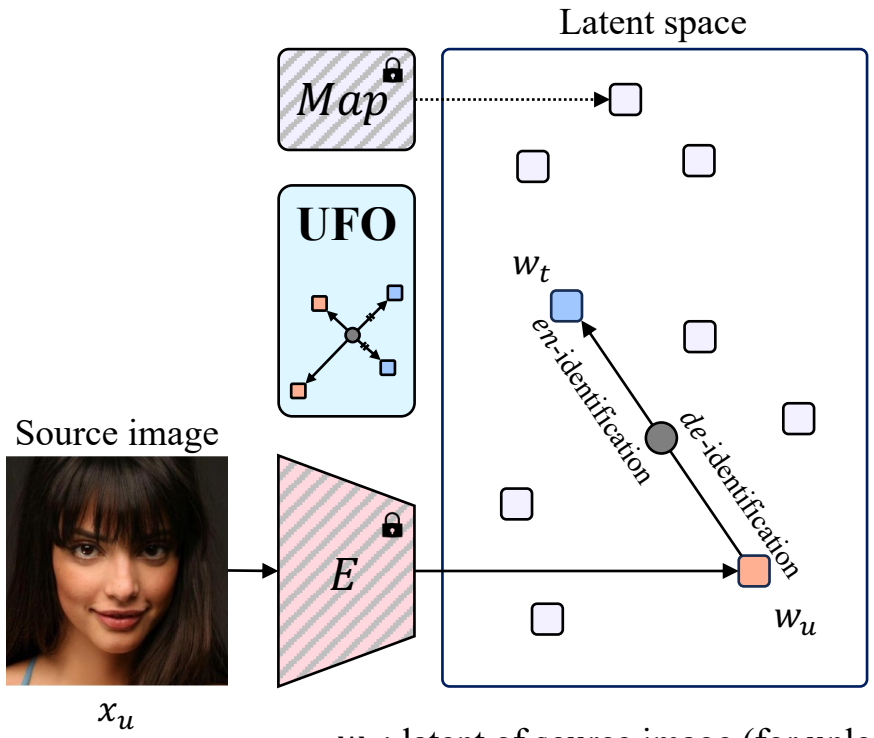


Anonymous identity (target for training)
High-quality generation

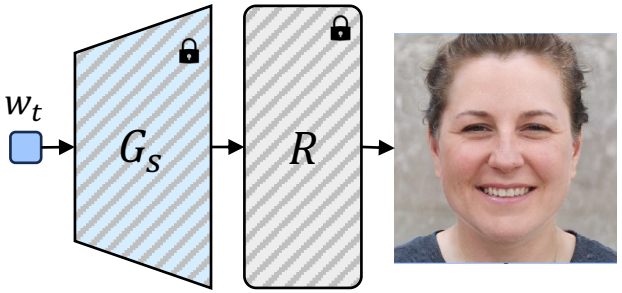
Identity removed (after training)
Preservation of high-quality generation

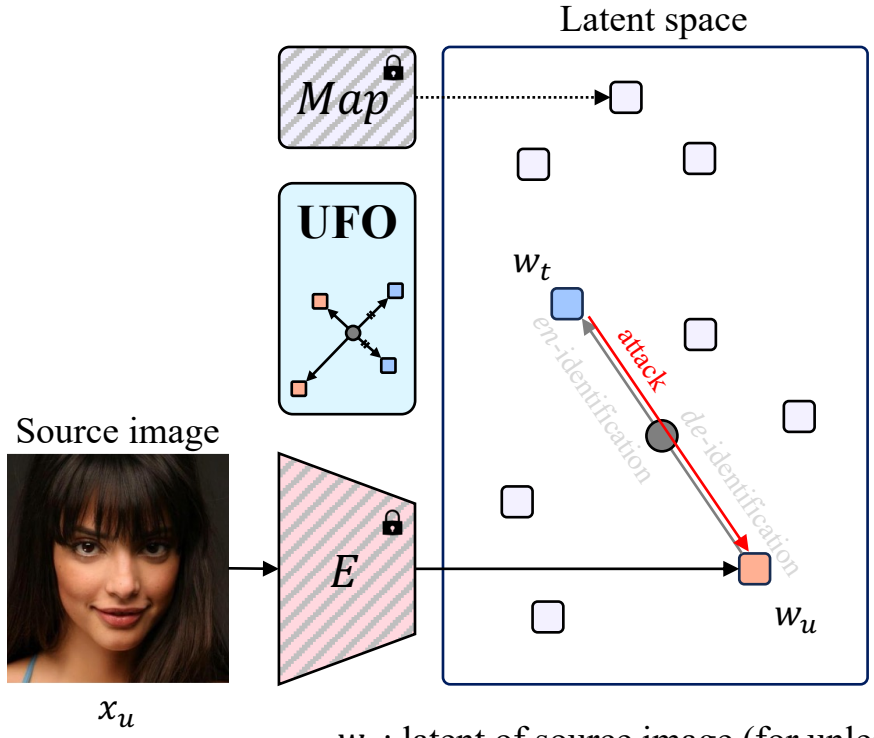
- Is the training process needed?
- What happens if we use the results generated for w_t without training G_u?

Answer

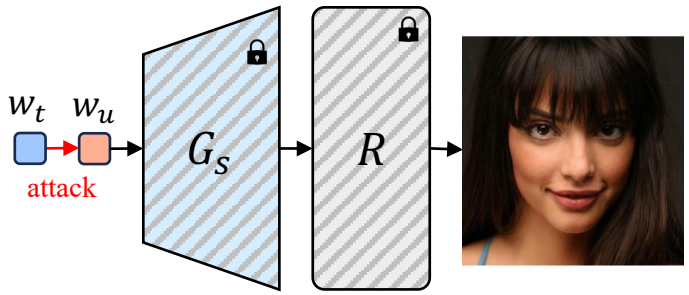


w_u: latent of source image (for unlearning)
w_t: target anonymized latent
▢ : known identities

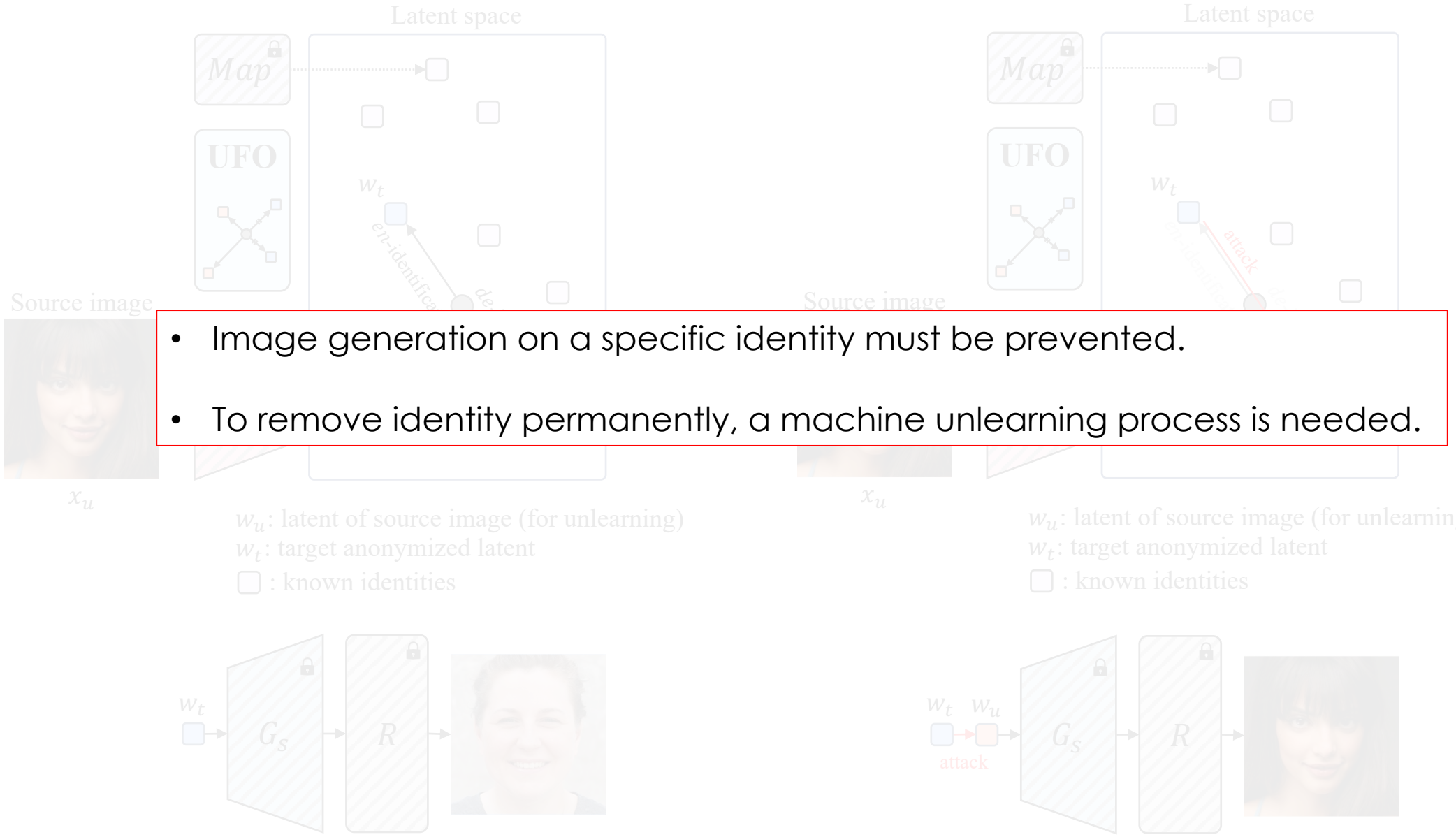




w_u: latent of source image (for unlearning)
w_t: target anonymized latent
▢ : known identities



Answer



- Image generation on a specific identity must be prevented.
- To remove identity permanently, a machine unlearning process is needed.

Details of GUIDE

Generative **U**nlearning for Any **IDE**ntity

Aim of GUIDE

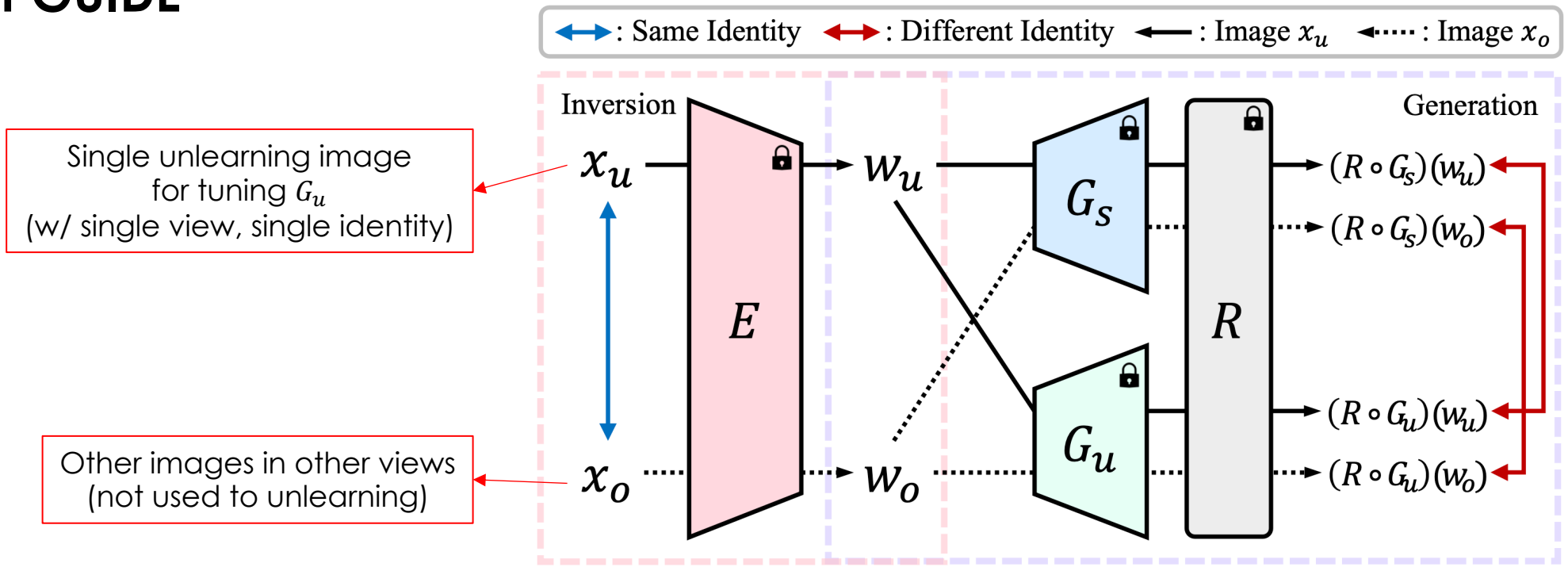


Figure 2. An illustration of *generative identity unlearning*. Upon GUIDE, the identity of the image generated from w_u, *i.e.*, inversion of the source image x_u by inversion network E, should exhibit a distinct identity when passed through the pre-trained generator G_s compared to the unlearned generator 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.

[StyleGAN] Tero Karras et al. "A style-based generator architecture for generative adversarial networks." CVPR 2019
[EG3D] Eric R. Chan et al. "Efficient geometry-aware 3d generative adversarial networks." CVPR 2022
[GOAE] Ziyang Yuan et al. "Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding." ICCV 2023

GUIDE

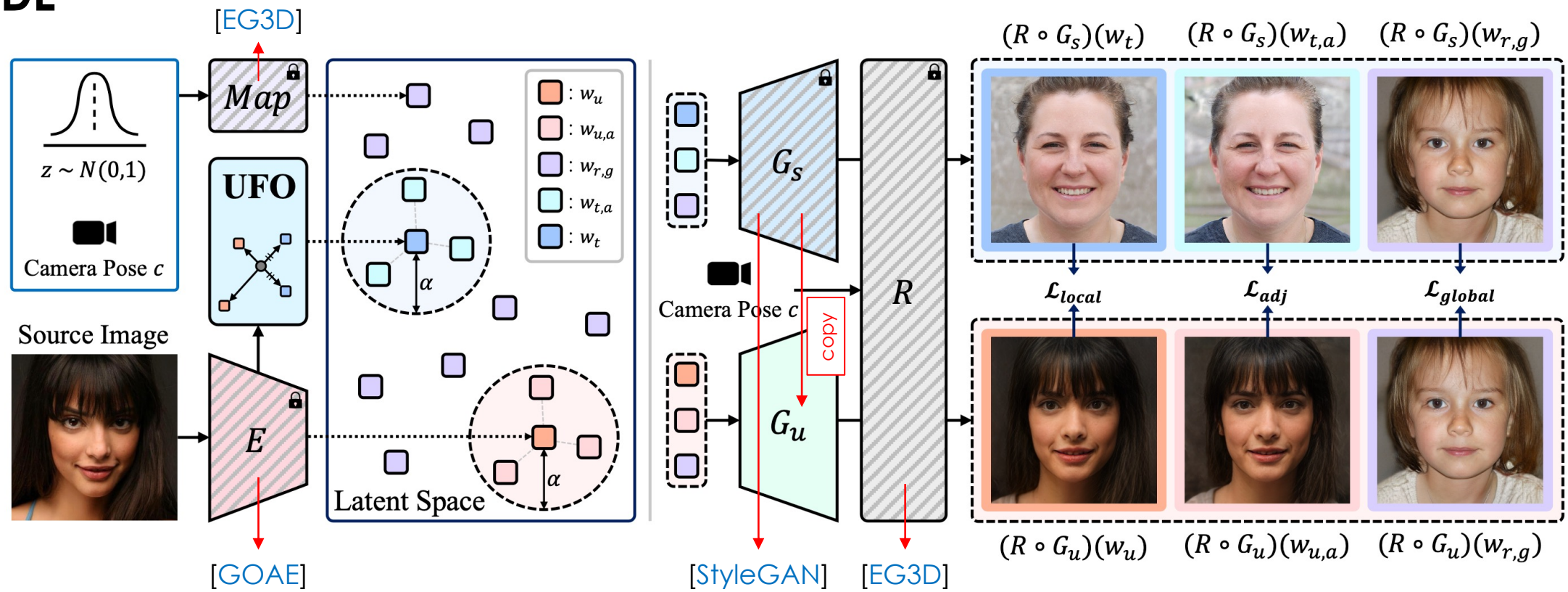


Figure 3. An overview of GUIDE. Starting with a source image, we employ a GAN inversion network E, specifically GOAE [52], to embed this image into the latent space of a pre-trained generative model, namely EG3D [4], obtaining the source latent code w_u. The target latent code w_t is designated through 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 $w_{r,g}$.

Components of GUIDE

Un-identifying Face On latent space (UFO, training set synthesis)

- de-identification process
- en-identification process

Latent Target Unlearning (LTU, loss terms)

- Local unlearning loss ($L_2, L_{per},$ and L_{id})
- Adjacency-aware unlearning loss ($L_2, L_{per}, and\ L_{id}$)
- Global preservation loss (L_{per} only)

UFO

$\bar{w}$: latent of mean human face
w_u: latent of source image (input for unlearning)
w_t: target anonymized latent (synthetic GT)

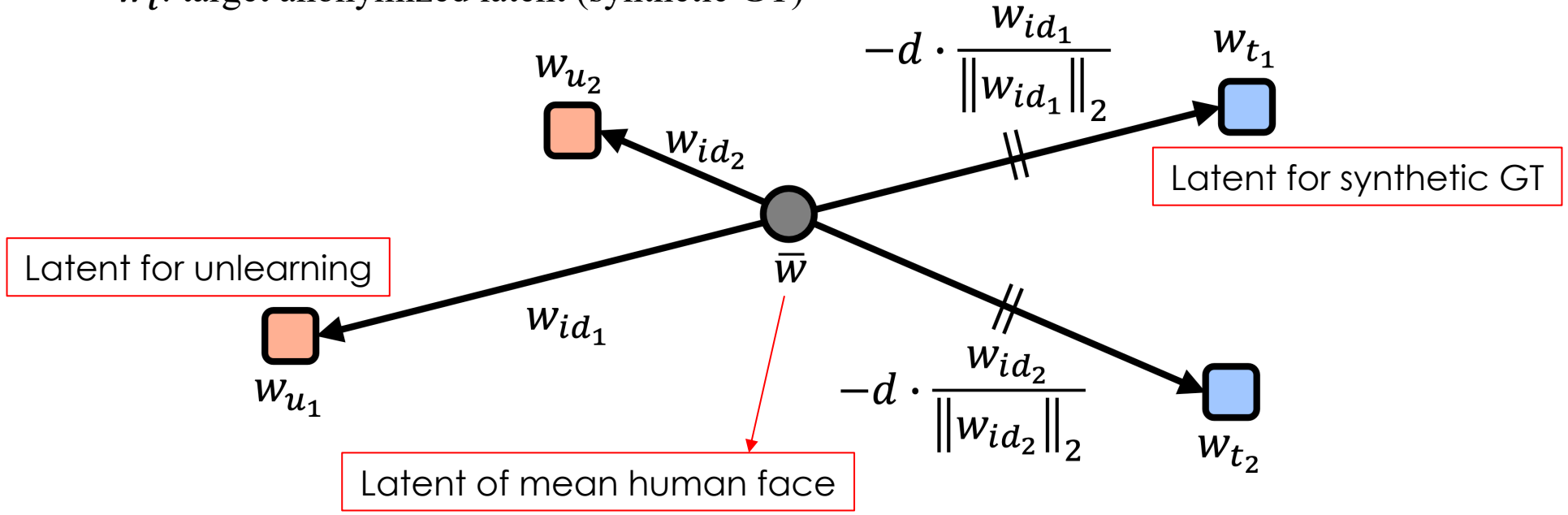
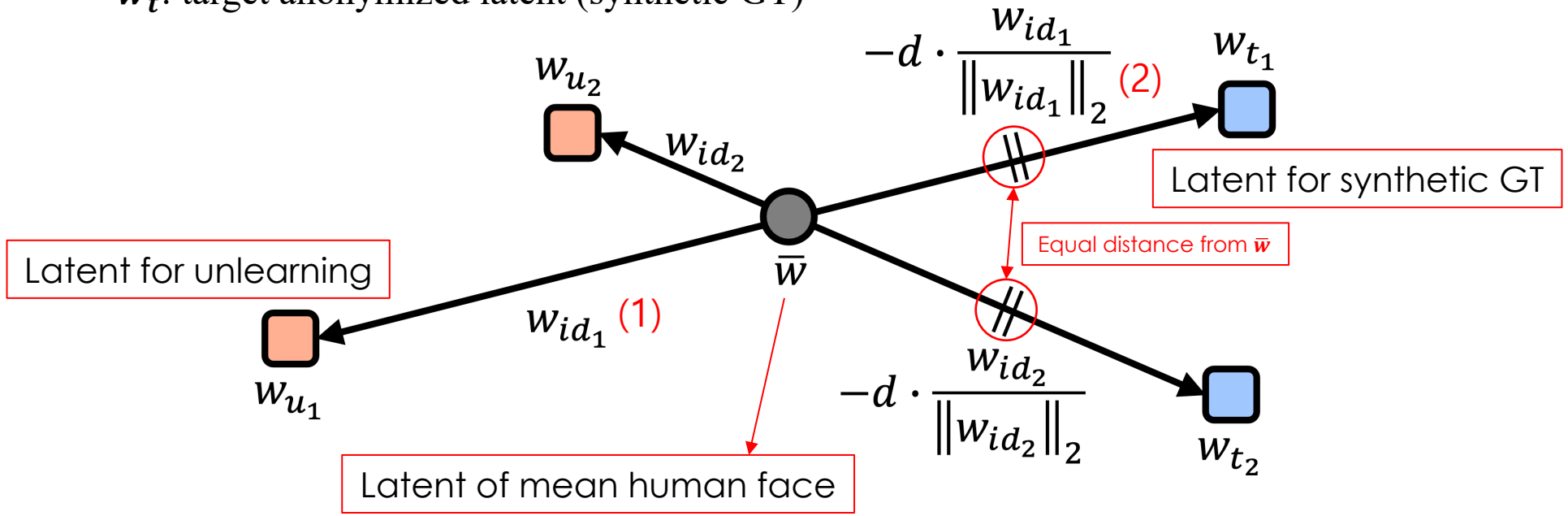


Figure 4. An illustration of Un-identifying Face On Latent Space (UFO). We define the identity of the source latent code by subtract it from the average latent code. We set the target latent code for our unlearning process by measuring an extrapolation between the source and average latent code with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

UFO

$\bar{w}$: latent of mean human face
w_u: latent of source image (input for unlearning)
w_t: target anonymized latent (synthetic GT)



(1) *de*-identification: compute identity vector of input source from mean human face $\bar{w}$

$$w_{id} = w_u - \bar{w}$$

(2) *en*-identification: generate opposite anonymous identity by extrapolation

$$w_t = \bar{w} - d\frac{w_{id}}{\|w_{id}\|_2} \quad \text{Eq (3)}$$

$-d$: negative magnitude (hyperparameter)

$\frac{w_{id}}{\|w_{id}\|_2}$: direction of identity

UFO

$\bar{w}$: latent of mean human face
w_u: latent of source image (input for unlearning)
w_t: target anonymized latent (synthetic GT)

Now, we can train generator G_u to minimize distance between x_u and x_t!

- x_u: unlearning sample
- x_t: target identity (anonymized identity)
- $x_u = G_u(w_u)$

(1) de-identification: compute identity vector of input source from mean human face $\bar{w}$

$$w_{id} = w_u - \bar{w}$$

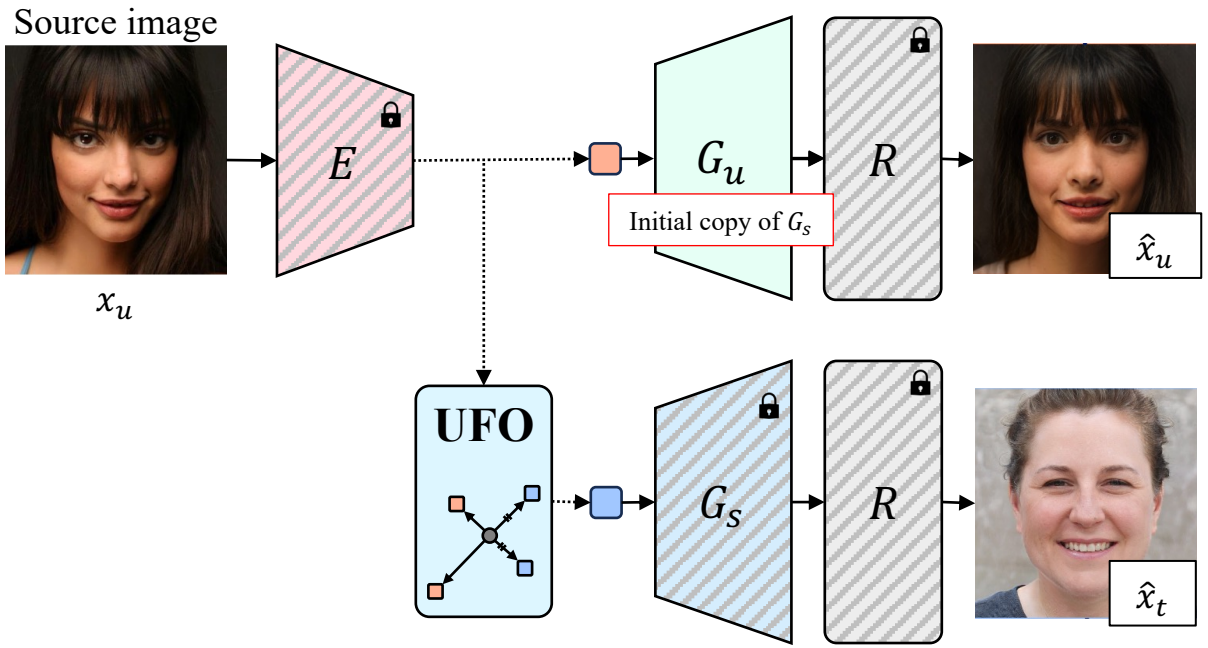
(2) en-identification: generate opposite anonymous identity by extrapolation

$$w_t = \bar{w} - d\frac{w_{id}}{\|w_{id}\|_2} \quad \text{Eq (3)}$$

$-d$: negative magnitude (hyperparameter)

$\frac{w_{id}}{\|w_{id}\|_2}$: direction of identity

LTU (1)
Local unlearning loss



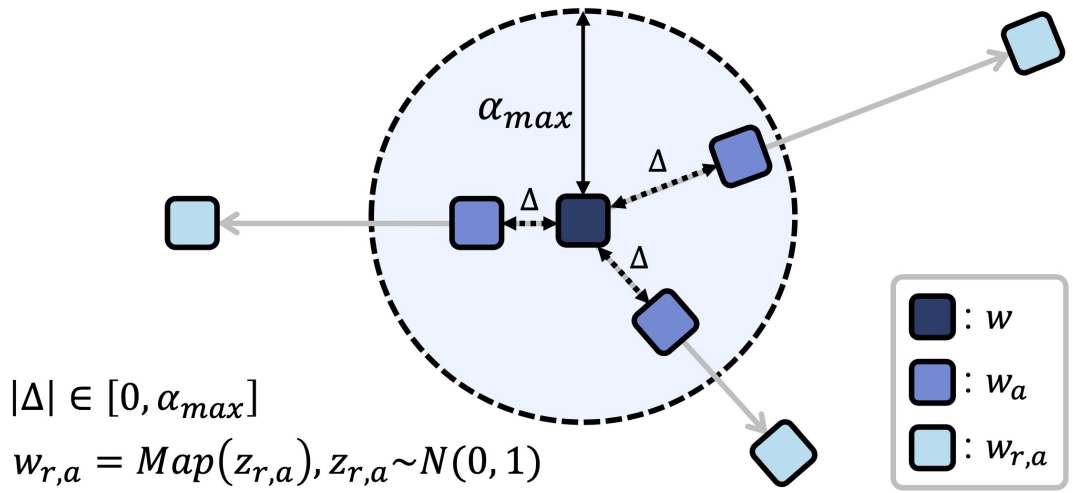
$F_u = G_u(w_u)$, generated feature map
$\hat{x}_u = R \circ G_u(w_u) = R \circ F_u$
Identity preservation (before training)

$F_t = G_s(w_t)$, generated feature map
$\hat{x}_t = R \circ G_s(w_t) = R \circ F_t$
Anonymous identity (target for training)

Pinch out a specific identity

$$\mathcal{L}_{local}(\hat{x}_u, \hat{x}_t) = \lambda_{L2}\mathcal{L}_{L2}(F_u, F_t) + \underbrace{\lambda_{per}\mathcal{L}_{per}(\hat{x}_u, \hat{x}_t)}_{\text{VGG feature distance}} + \underbrace{\lambda_{id}\mathcal{L}_{id}(\hat{x}_u, \hat{x}_t)}_{\text{ArcFace}}. \quad (4)$$

LTU (2)
Adjacency-aware unlearning loss



1) Set latent w_u for unlearning
2) Generate random (r) adjacency (a) latent
   - Sample noise vector $z_{r,a}$
   - Map $z_{r,a}$ into latent space via mapping network
   - Get latent $w_{r,a}$
3) Generate vector Δ to jitter unlearning latent by Eq (5)
   - To regularize the adjacency latent into boundary
4) Apply Δ for both w_u and w_t
   - $w_{u,a} = w_u + \Delta$
   - $w_{t,a} = w_t + \Delta$
5) Repeat N_a times to create an adjacency set

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 $w_a = w + \Delta$.

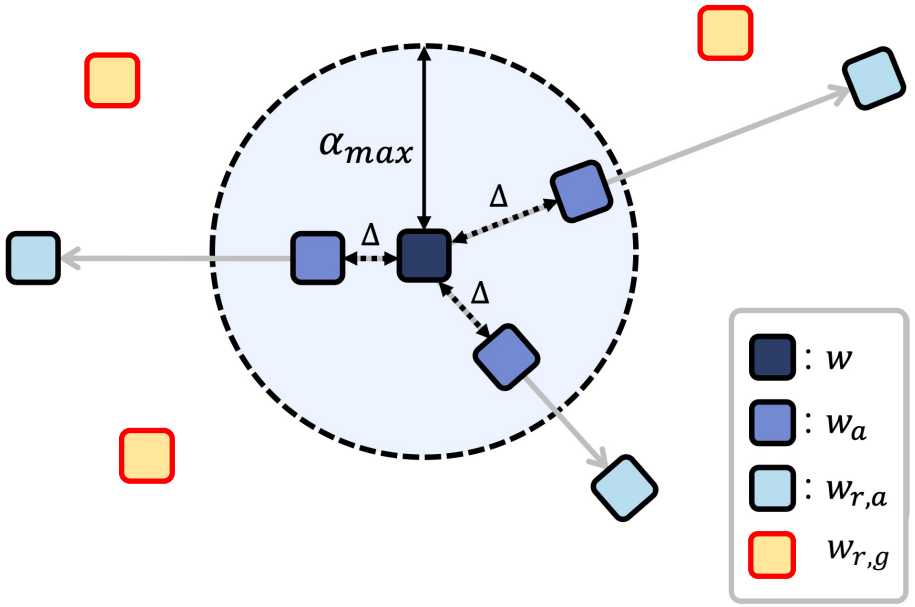
$$\Delta = \{\alpha^i \cdot \frac{w^i_{r,a} - w_u}{\|w^i_{r,a} - w_u\|_2}\}^{N_a}_{i=1}, \tag{5}$$

Remove vicinity (neighborhood) identities

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

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

LTU (3)
Global preservation unlearning loss



Preserve high-quality generation
To preserve the generation quality of G_u for other identities after unlearning.

1) Set latent w_u for unlearning
2) Generate random (r) global (g) latent
   - Sample noise vector $z_{r,g}$
   - Map $z_{r,g}$ into latent space via mapping network
   - Get latent $w_{r,g}$
3) Repeat N_g times to create an adjacency set

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

Experiments

Methods to comparison

GUIDE (baseline)

- Set target latent as the average of total unlearning set (except single unlearning input)
- $w_t = \frac{1}{n}\sum_{i=1}^{N} w_u^i$

GUIDE (w/ UFO)

- Full model of GUIDE
- $w_t = \bar{w} - d\frac{w_{id}}{\|w_{id}\|_2}$

Qualitative results (1)

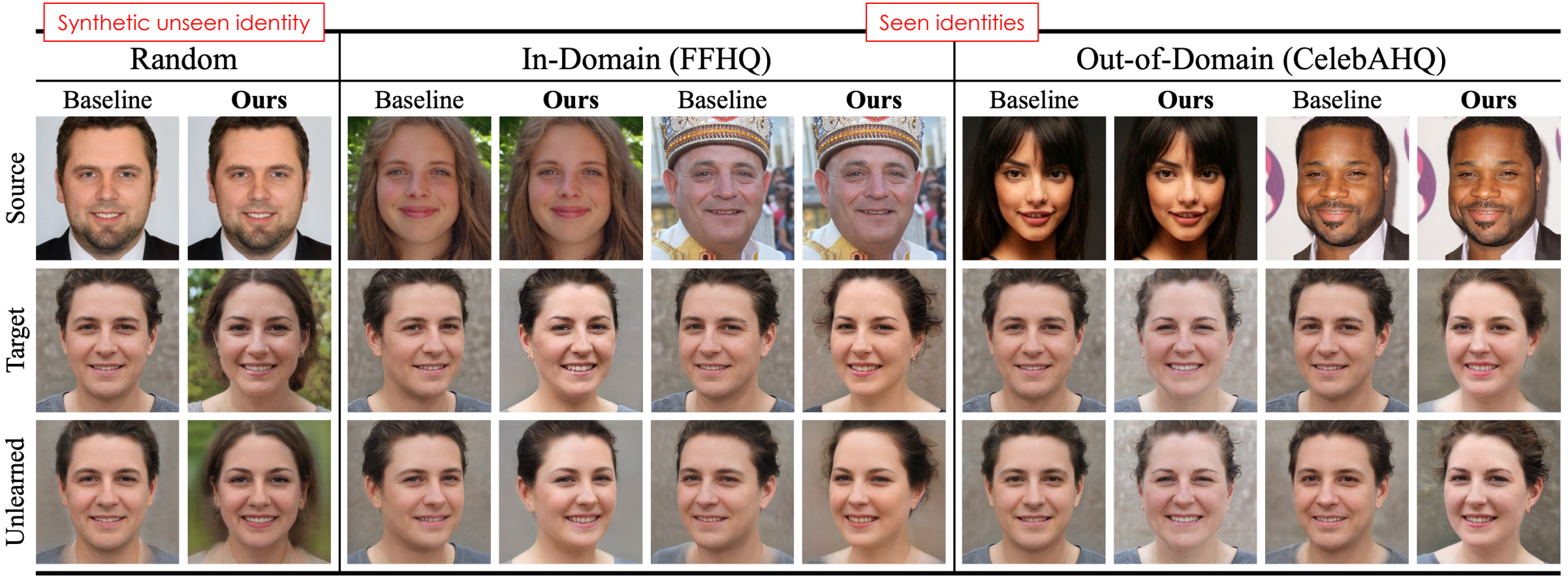


Figure 6. Qualitative results of GUIDE and the baseline in generative identity unlearning task. For the given source image each (the first row), GUIDE and the baseline tried to erase the identity in the pre-trained generator. The images in the second and third row are the target and unlearned images,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

Qualitative results (2)

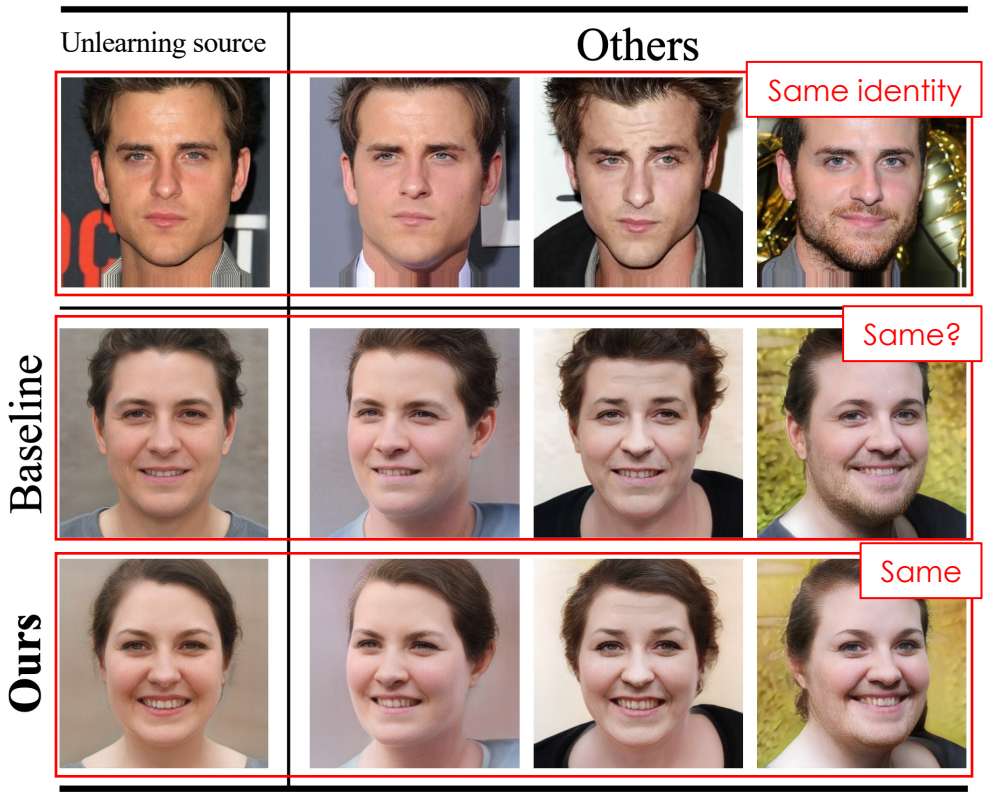


Figure 7. Qualitative results of GUIDE and the baseline on a multi-image test using CelebAHQ dataset. We additionally utilized images that are unseen during unlearning, to show how thoroughly erase the given identity.

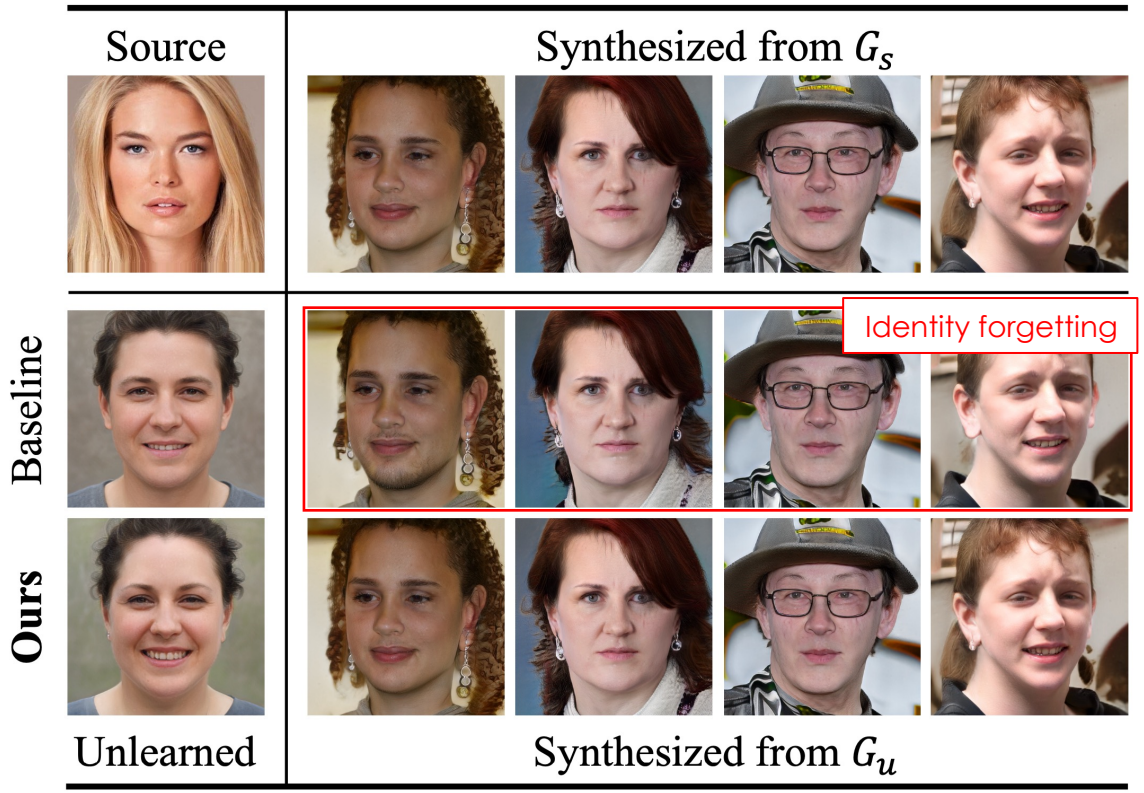


Figure 8. Qualitative comparison between GUIDE and the baseline on the preservation of the generation quality of other identities. GUIDE generates images almost identical to those synthesized by G_s, whereas the baseline often results in noticeable changes, *e.g.*, beard shape, hairstyle change, hat.

[ID] Yuge Huang et al. "Curricularface: adaptive curriculum learning loss for deep face recognition." CVPR 2020
[FID] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." NIPS 2017

Quantitative results (1)

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.

- ID: ID similarity of unlearned identities
  - CuricularFace, face recognition network (results identity embedding vector)
- FID_{pre}: distribution shift (generation quality) of G_u compared to pre-trained generator G_s
  - $FID_{pre} = FID(R \circ G_u(w_u), R \circ G_s(w))$
- $\Delta\text{FID}_{\text{real}}$: distribution shift (generation quality) of G_u compared to real images (FFHQ)
  - $\Delta FID_{real} = FID(R \circ G_u(w_u), x_{FFHQ})$

Quantitative results (2)

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.

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.

$\text{ID}_{\text{others}}$: ID similarity of unlearned identities

Ablation studies

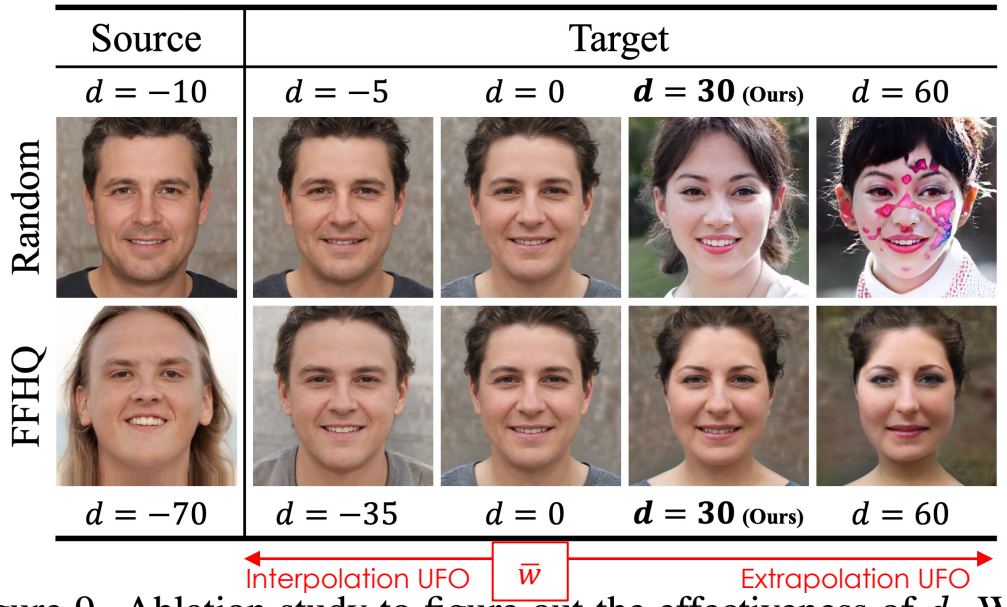


Figure 9. Ablation study to figure out the effectiveness of d. We visualized target images corresponding to each source image with different values of d. The target images were generated using target latent codes derived from interpolated latent codes, the average latent code ($d = 0$), or extrapolated 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	w/o vicinity removal
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	vicinity removal (wide)

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.

$\mathcal{L}_{local}$	$\mathcal{L}_{global}$	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.

Supplementary Material

Why adjacency-aware unlearning loss?

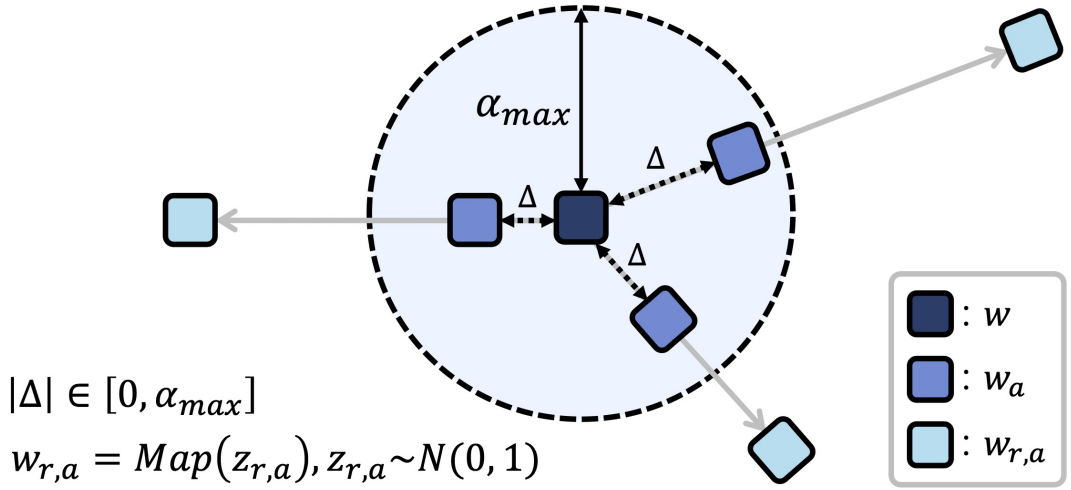


Figure 5. An illustration of determining latent codes near a latent code w in adjacency-aware unlearning loss. We first sample a latent code $w_{r,a}$ which is derived from a random noise vector $z_{r,a}$ via the mapping network $Map(\cdot)$, *i.e.* $w_{r,a} = Map(z_{r,a})$. Next, we compute the direction 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 $w_a = w + \Delta$.

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

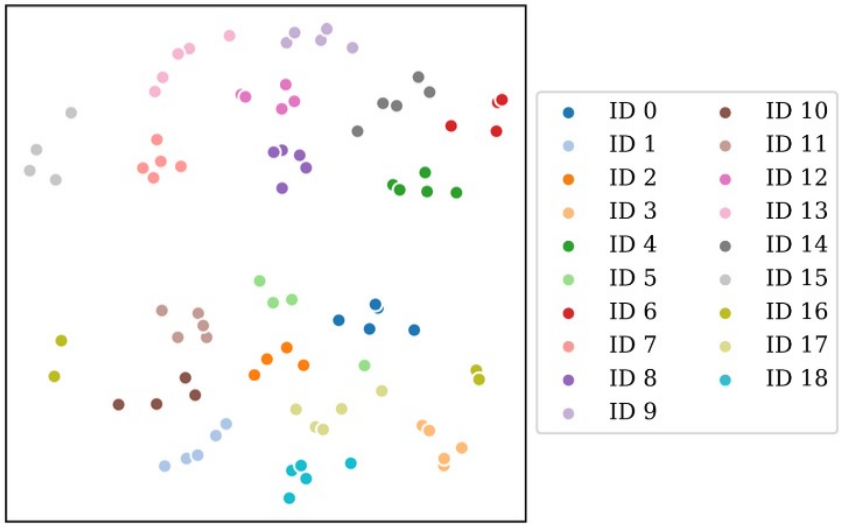


Figure 2. The relationship between the images and their identities in the latent space with t-SNE [6]. Points of the same color denote the same identity. We used 5 images per identity from CelebAHQ dataset.

Why adjacency-aware unlearning loss?

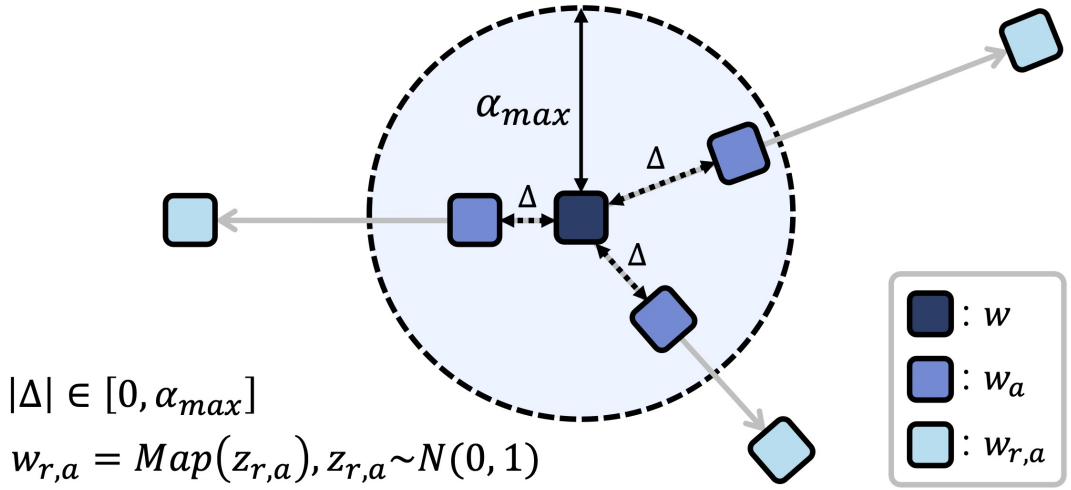


Figure 5. An illustration of determining latent codes near a latent code w in adjacency-aware unlearning loss. We first sample a latent code $w_{r,a}$ which is derived from a random noise vector $z_{r,a}$ via the mapping network $Map(\cdot)$, *i.e.* $w_{r,a} = Map(z_{r,a})$. Next, we compute the direction 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 $w_a = w + \Delta$.

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

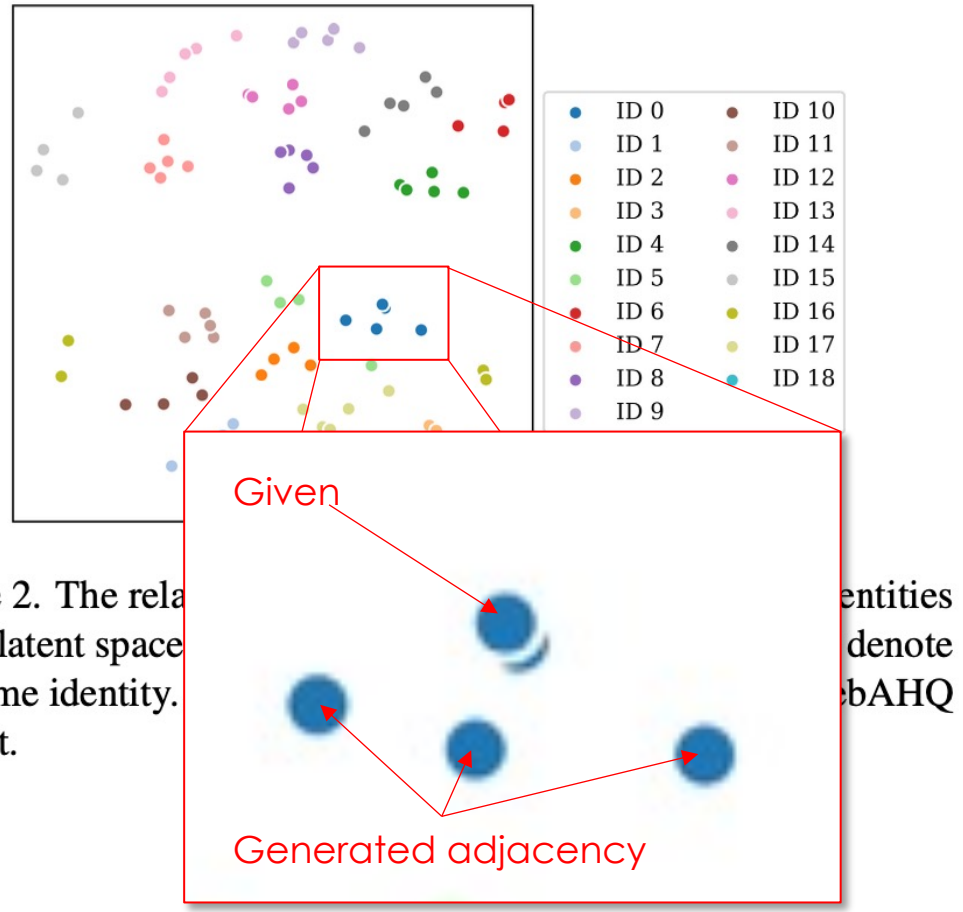


Figure 2. The rela ntities in the latent space denote the same identity. bAHQ dataset.

Allows to unlearn specific identity with only a single image!

Results

Figure 6. Additional qualitative results with CelebAHQ dataset.

Is GUIDE also working for another domain?

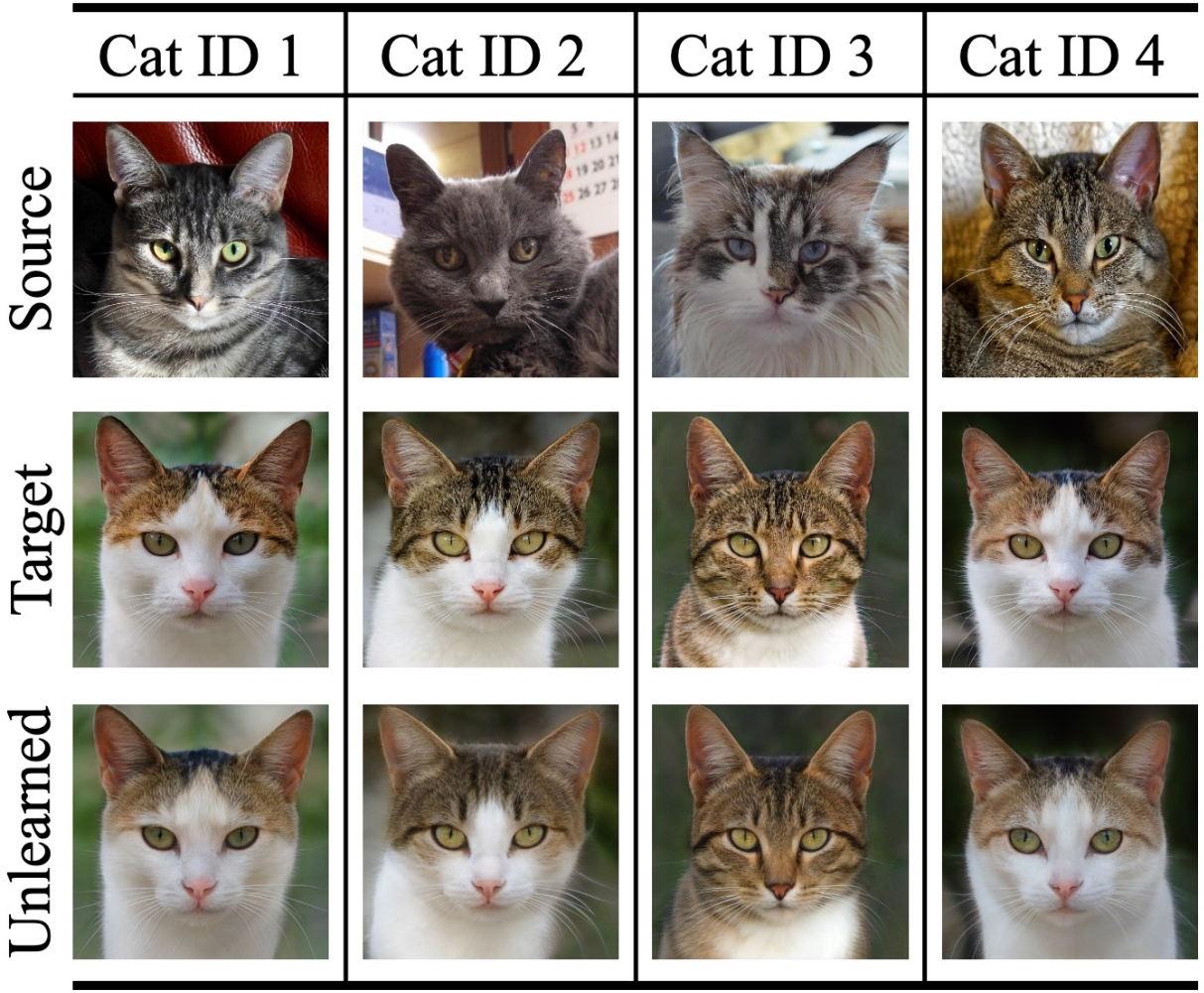


Figure 3. Qualitative results of generative identity unlearning on AFHQv2-Cat dataset.

* SG2: StyleGan V2

Is GUIDE also working for another generative model?

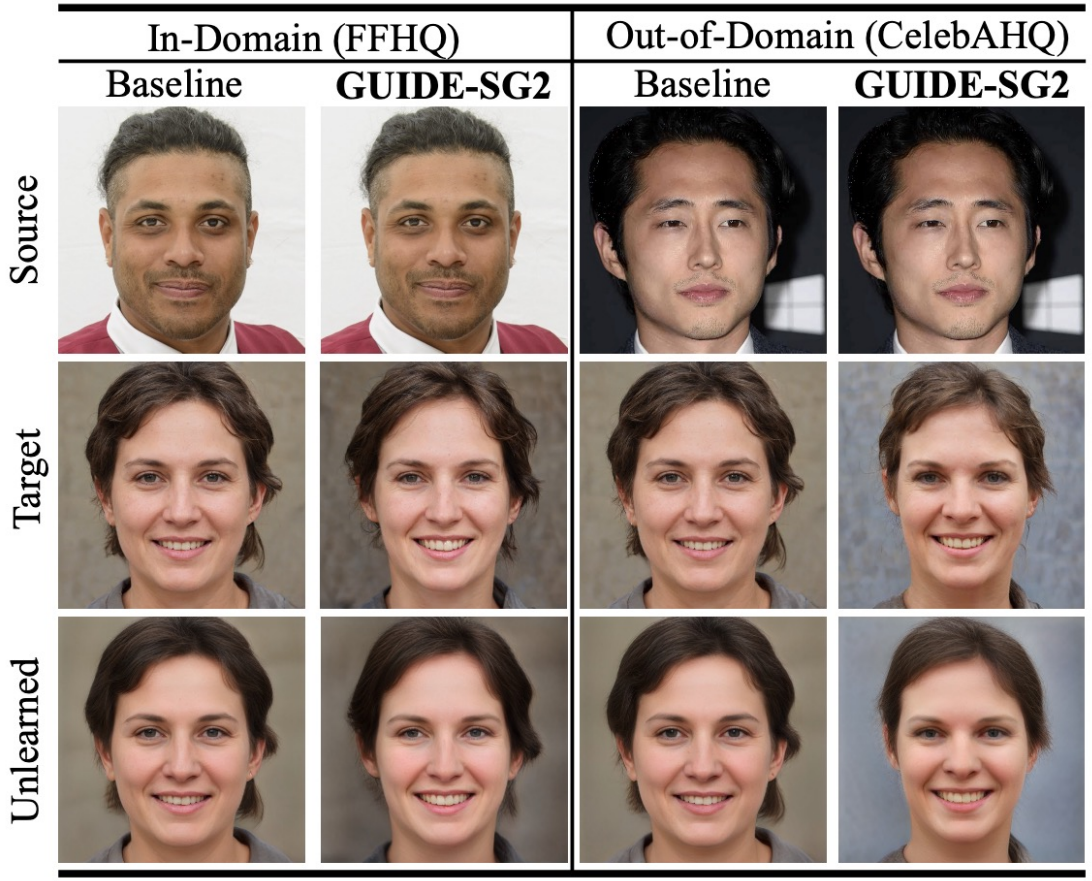


Figure 8. Qualitative results of GUIDE-SG2 and the baseline. For the given source image each (the first row), GUIDE-SG2 and the baseline tried to erase the identity in the pre-trained generator. The result are shown on the second row. Images in the third row are the target image in our unlearning process.