ROVER: Robust Generative Continual Identity Unlearning Against Relearning Attacks

Tairan Huang[1*], Qiang Chen[2*], Beibei Hu[3], Yunlong Zhao[4], Hongyan Xu[2†], Zhiyuan Chen[5], Yi Chen[6], Xiu Su[2†]

[1]School of Computer Science and Engineering, Central South University, Changsha, China
[2]Big Data Institute, Central South University, Changsha, China
[3]College of Textile and Fashion, Hunan Institute of Engineering, Changsha, China
[4]School of Electronics and Information Engineering, Central South University, Changsha, China
[5]Institute of Data Science, University of Hong Kong, Hong Kong, China
[6]School of Engineering, HKUST, Hong Kong, China
{tairanhuang, xiusu1994, zhaoyl741}@csu.edu.cn, {qiangchen.sh, mindyhbb}@gmail.com, hongyan.xu@unsw.edu.au, this@zyc.ai, yichen@ust.hk

Abstract

Recent generative unlearning models synthesize high quality samples while protecting private information by unlearning the identity. However, existing generative identity unlearning methods face two challenges in multi-identity unlearning: 1) identity conflicts, which cause conflicts of model parameters in the continuous erasure of multiple identities; 2) fragile unlearning, where the model's unlearning ability deteriorates or fails under malicious attacks. In this paper, we introduce a critical yet under-explored task called robust multi-identity unlearning, with the goals of resolving identity conflicts to achieve interference-free unlearning and protecting against malicious attacks to achieve robust unlearning. To satisfy these goals, we propose a novel framework, **RO**bust generati**VE** continual identity unlearning against **R**elearning attacks (ROVER). By filtering unlearning requests with latent similarity, our method effectively isolates benign unlearning from malicious attacks to preserve identity removal integrity. Meanwhile, residual orthogonal resonator resolves identity conflicts in the continuous erasure of multiple identities, preserving stability in benign continual unlearning. Moreover, we introduce the phantom guard network to block malicious attacks by absorbing adversarial gradients, ensuring irreversible identity unlearning. The extensive experiments demonstrate that our proposed method achieves state-of-the-art performance on the task of robust multi-identity unlearning against relearning attacks.

Introduction

Recently, Generative Adversarial Networks (GANs) (Wang et al. 2025; Lee and Min 2025) have achieved remarkable success in producing high-quality and realistic images in a wide range of domains from artistic image generation (Chu et al. 2024; Saxena et al. 2024) to realistic face synthesis (Chen et al. 2024; Zhang, Guo, and Zhou 2024). These advancements are largely attributed to improvements in architectural design and the availability of large-scale training datasets that enable generators to capture intricate visual details. However, these advancements also raise significant privacy concerns (Huang et al. 2025), particularly due to the potential misuse in synthesizing images of real individual identities or amplifying harmful concepts. For example, Deepfakes (Xu et al. 2023; Yan et al. 2023) builds realistic images or videos of people after learning sufficiently from existing data to raise major ethical and privacy concerns.

Generative unlearning (Sun et al. 2025; Zhang et al. 2024b) involves selectively modifying pre-trained generative models to make them forget sensitive data to protect personal privacy (Zhang et al. 2024a; Gandikota et al. 2023), including concept unlearning and identity unlearning. Recently, several diffusion-based generative unlearning methods (Zhang et al. 2024c; Huang et al. 2024a; Moon, Cho, and Kim 2024) incorporate editing or regularization techniques to selectively remove semantic classes or concepts during the generation process. In contrast, GUIDE (Seo et al. 2024) is proposed as the first effort to image identity unlearning in generative models, achieving the erasure of the entire identity with a single image from the generator while preserving the pre-trained model's ability to generate other identities. Compared to generative concept unlearning, generative identity unlearning remains largely unexplored.

Despite its novelty, GUIDE still exhibits two key limitations when applied to multi-identity unlearning. First, the continual requests to unlearn multiple identities can generate conflicting gradient signals, causing previously unlearned identities to re-emerge or degrading overall generation quality. Each request the generator is asked to unlearn a new identity, its parameters are updated to suppress that specific image. However, these updates can interfere with or even reverse the effects of earlier unlearning requests. Second, continual requests to unlearn multiple identities can lead to unlearning invalidity when subjected to malicious attacks such as relearning attacks. Malicious users can easily restore unlearned identities by fine-tuning the model with images that closely resemble the original source images, thereby offsetting the effects of unlearning. Without dedicated defenses, standard fine-tuning offers no resistance to such reversions,

*These authors contributed equally.
†Corresponding author.

making unlearning fragile against malicious attacks.

In this work, we propose a novel framework, RObust generatiVE continual identity unlearning against Relearning attacks, named ROVER. ROVER consists of three crucial components that effectively erase the multiple identities in continual unlearning requests and completely protect against relearning attacks. Specifically, the *Similarity Navigator* computes latent similarity scores to filter unlearning requests, thereby separating benign unlearning from relearning attacks and preserving the integrity of erased identities. The *Residual Orthogonal Resonator* applies orthogonal residual constraints to resolve conflicts among multiple identities during sequential unlearning, ensuring stable and accurate removal of each identity. The *Phantom Guard Network* redirects adversarial gradients into a dedicated phantom branch, effectively blocking relearning attacks and guaranteeing irreversible unlearning of targeted identities. The comprehensive experiments in three scenarios show that ROVER can successfully remove the multiple identities of source images from a pre-trained generative model and eliminate the effects of relearning attacks to achieve state-of-the-art (SOTA) performance.

Our contributions are summarized as follows:

- For the first time, we propose a novel task, robust multi-identity unlearning, which aims to achieve interference-free unlearning and protect against malicious attacks.
- We propose the similarity navigator to efficiently filter benign unlearning requests and malicious attacks, combined with the residual orthogonal resonator and phantom guard network to achieve stable and robust unlearning for multiple identities.
- We conduct extensive experiments on three scenarios to demonstrate the effectiveness and robustness of ROVER, achieving state-of-the-art performance on the task of robust multi-identity unlearning.

Related Work

Generative Unlearning

Recent advances in generative unlearning have sought to equip powerful generative models with the ability to selectively unlearn specific data or concepts (Tiwary et al. 2023; Qiang Chen 2025; Fan et al. 2024; Panda and Prathosh 2024). Existing works are primarily built around diffusion models (Gao et al. 2024; Dai and Gifford 2023; Wu et al. 2025, 2024; Gandikota et al. 2023). Moreover, Text-to-image and text-to-video settings have also gained attention. Liu et al. (Liu and Tan 2024) and Kumari et al. (Kumari et al. 2023) extend the unlearning capabilities to multimodal diffusion models, targeting specific visual semantic correspondences. Although diffusion-based methods can remove concepts, they cannot prevent models from generating specific identities (Seo et al. 2024). Beyond diffusion models, GUIDE (Seo et al. 2024) presents the first generative identity unlearning method based on GANs, that effectively removes the identity from a single image. Compared to GUIDE, our work introduces orthogonal residual constraints to achieve effective erasure of multiple identities under continual unlearning requests.

Continual Learning

Continual learning for generative models has attracted significant attention, as it enables models to assimilate new data distributions without catastrophic forgetting. Early efforts (Zhai et al. 2019; Cong et al. 2020) introduce memory replay and distillation to retain previously learned conditional mappings. Hyper-LifelongGAN (Zhai, Chen, and Mori 2021), OCM (Ye and Bors 2022a), DEGM (Ye and Bors 2022b), and KFC (Huang et al. 2024b) explore dynamic expansion and cooperative memorization, which scales continual generation via progressive network growth. Besides, Forget-Me-Not (Zhang et al. 2024a) and Selective Amnesia (Heng and Soh 2023) serve as continual learning methods to selectively erase undesired concepts while preserving overall performance. However, the above methods mainly aim to mitigate catastrophic forgetting by retaining previously acquired knowledge or deleting broad concepts. In contrast, we focus on continuous erasure for specific identities, being the first continual identity unlearning method in generative models.

Jailbreak Attacks

Jailbreak attacks against large-scale generative models have emerged as a critical security concern, forcing models can be coerced to produce prohibited or sensitive content (Zhang et al. 2020; Xie et al. 2021; Yang et al. 2017). Early attack strategies exploited prompt injection (Hu and Pang 2021; He et al. 2022), in which adversaries craft innocuous-looking inputs that surreptitiously override system instructions or content policies. Some studies have focused on designing hand-crafted adversarial prompts (Yong, Menghini, and Bach 2023; Li et al. 2025), while others have explored automatically generating prompts through gradient-based optimization (Jones et al. 2023; Li, Bradshaw, and Sharma 2019) or genetic method (Zhou et al. 2022; Wu et al. 2021). However, they primarily focus on the introduction of harmful behaviors. In this work, we introduce the phantom generator to prevent malicious attacks from restoring the model's unlearning ability while maintaining the integrity of the unlearned generator.

Methodology

Preliminaries

Generative Continual Identity Unlearning. The set of source images is defined as $\mathcal{S} = \{s_n\}_{n=1}^{N}$ whose identities x_n are to be unlearned and the set of target images is defined as $\mathcal{T} = \{t_n\}_{n=1}^{N}$, where N denotes the number of available source and target images. Each image is mapped into the latent space to obtain the latent code z_{s_n} and z_{t_n} by a frozen inversion network E (Yuan et al. 2023):

$$z_{s_n} = E(s_n), \quad z_{t_n} = E(t_n). \tag{1}$$

Let $G_s(\cdot;\theta_s)$ denote the pre-trained generator EG3D (Chan et al. 2022) with fixed parameters θ_s and $G_u^i(\cdot;\theta_u^i)$ denote the unlearned generator at request i with parameters θ_u^i. The neural renderer is defined as $R(\cdot)$ with a super-resolution (Karras, Laine, and Aila 2019). At continual unlearning request i, we use the images n from S and $\mathcal{T}$ for training. The

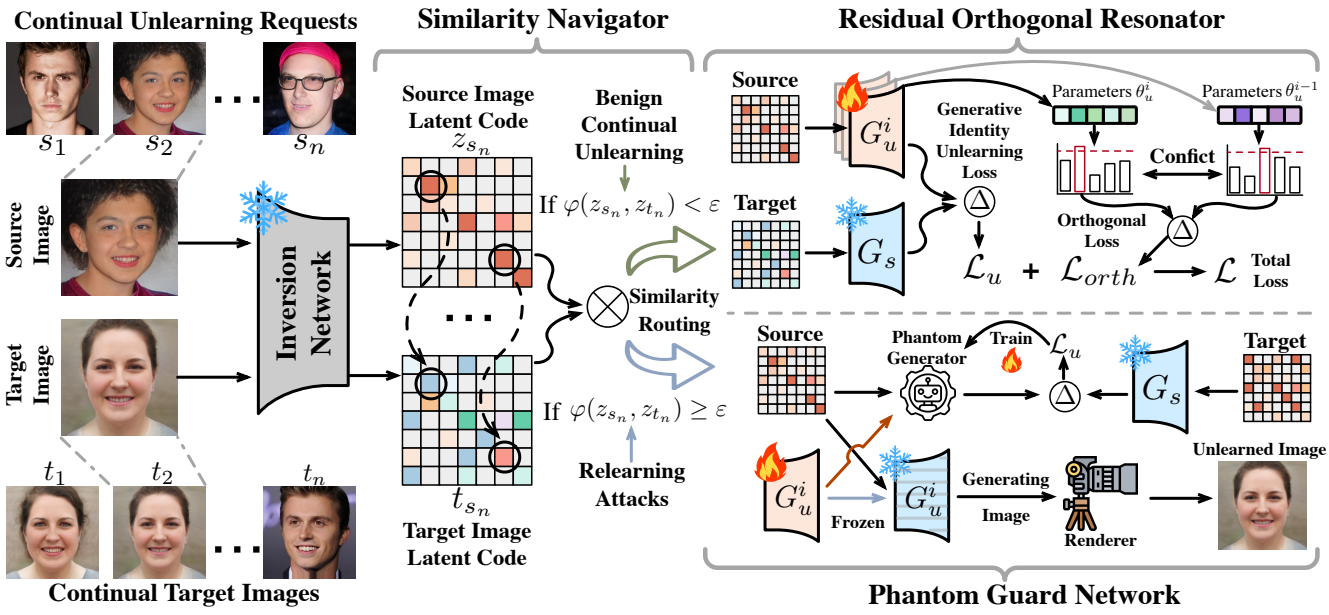


Figure 1: The overall framework of ROVER.

reconstructed images of s_n and t_n can be expressed as:

$$\hat{s}_n = R(G_u^i(z_{s_n}; \theta_u^i)), \hat{t}_n = R(G_s(z_{t_n}; \theta_s)). \quad (2)$$

The goal is to continuously update the unlearned generator G_u^i to transform each source image s_n with the identity x_n into the identity corresponding to the target image t_n, achieving the unlearning of the original identity. The unlearned generator is continually optimized through the generative identity unlearning loss $\mathcal{L}_u$ (Seo et al. 2024):

$$\mathcal{L}_u = \underbrace{\mathcal{L}_e(\hat{s}_n, \hat{t}_n)}_{\text{Unlearn}} + \underbrace{\mathcal{L}_r(r_n)}_{\text{Retain}}, \quad (3)$$

$$\theta_u^i = \arg\min_{\theta} \mathcal{L}_u^i(\cdot; \theta) \quad \text{s.t.} \quad \theta_u^i\big|_{\theta=\theta_u^{i-1}}, \quad (4)$$

with initialization $\theta_u^0 = \theta_s$. $\mathcal{L}_e$ and $\mathcal{L}_r$ represent the unlearning loss and retaining loss (Seo et al. 2024), respectively, and r_n represents the image in retained set D_r. The details are provided in the appendix. This ensures that each new unlearning request retains all previously acquired the unlearning capabilities.

Relearning Attacks. In generative continual identity unlearning, processing unlearning requests sequentially can easily lead to the target of one request being overwritten or relearned by subsequent requests. We perform a relearning attack by fine-tuning θ_u on a small relearning attack set $A = \{a_m\}_{m=1}^M$, where M denotes the number of available relearning images. The identity distribution of this attack set is very close to that of the source image set. Under relearning attacks, the target image is replaced with a randomly selected image from the relearning attack set. The goal of the relearning attack is to restore the image identity that is removed by unlearning, defined as:

$$\theta_u^{i'} = \arg\min_{\theta} \mathcal{L}_u^{i'}(\hat{s}_n, \hat{t}_m; \theta), \quad (5)$$

where $\theta_u^{i'}$ is the parameter for unlearned generator after the relearning attack and $\hat{t}_m = R(G_s(E(a_m); \theta_s))$ is the reconstructed image from a_m. The relearning attacks result in high identity similarity between the reconstructed images and the source images, reversing the unlearning effect.

Similarity Navigator

It is crucial to accurately distinguish between benign unlearning requests and malicious relearning attacks to deploy targeted defenses and prevent the leakage of sensitive identity information. To robustly distinguish between benign continual unlearning and relearning attacks, we propose *Similarity Navigator* that monitors the cosine similarity between each source–target latent pair. Note that the relearning attack replaces the target image set, resulting in close direct distances between source–target latent pairs with high similarity. For the n-th source image s_n and its guiding target image t_n under the unlearning request i, we compute:

$$\varphi_n^i = \frac{\langle z_{s_n}, z_{t_n} \rangle}{\|z_{s_n}\| \, \|z_{t_n}\|}, \quad z_{s_n} = E(s_n), \; z_{t_n} = E(t_n), \quad (6)$$

at continual request i. Intuitively, small values of φ_n^i indicate that the current update is driven by genuine unlearning signals, whereas large values signal a potential relearning attack. We introduce a fixed threshold ε as the decision boundary between these two regimes. Chosen via cross-validation or prior heuristic, ε reflects the maximum allowable similarity under normal continual unlearning. By fixing ε, the *Similarity Navigator* remains simple to tune and deploy, yet effectively adapts to varying data distributions. At each request, the navigator routes the gradient $\Delta\theta_u^i$ according to:

$$\text{Route}(\Delta\theta_u^i) = \begin{cases} \text{Unlearning Path}, & \varphi_n^i < \varepsilon, \\ \text{Relearning Path}, & \varphi_n^i \geq \varepsilon, \end{cases} \quad (7)$$

where the unlearning path routes to the proposed *Residual Orthogonal Resonator* and the relearning path routes to the proposed *Phantom Guard Network*, respectively. This lightweight mechanism incurs only one additional cosine-similarity evaluation. By consistently enforcing this routing logic, the *Similarity Navigator* safeguards the continual unlearning against relearning attacks, ensuring robust and reliable multi-identity unlearning.

Residual Orthogonal Resonator

The generative continual identity unlearning presents a unique challenge in that the model must sequentially erase the identity of each new source image without inadvertently restoring or corrupting those identities already unlearned. The fine-tuning of the generator parameters often produces update directions that conflict across continual unlearning requests, leading to catastrophic interference. To address this conflict, we introduce the *Residual Orthogonal Resonator* (O-Res), a simple yet powerful module that enforces orthogonality between successive parameter updates, thereby preserving all previously acquired unlearning signals. Let θ_u^{i-1} denote the generator's parameters after the previous unlearning request and θ_u^i represent the parameters obtained by the unlearning loss $\mathcal{L}_u$ at request i. We define the residual update at request i as:

$$\Delta\theta_u^i = \theta_u^i - \theta_u^{i-1}, \tag{8}$$

which captures the net change introduced by the current unlearning request. To prevent the current unlearning from overwriting previous unlearning requests, O-Res computes an orthogonal loss by projecting this residual onto the previous parameter vector and measuring the orthogonal component. We can express the orthogonal loss directly in terms of $\Delta\theta_u^i$ as:

$$\mathcal{L}_{orth}^i = \left\| \Delta\theta_u^i - \frac{(\theta_u^{i-1})^\top \Delta\theta_u^i}{\|\theta_u^{i-1}\|^2}\, \theta_u^{i-1} \right\|^2. \tag{9}$$

This loss exactly measures the squared norm of the component of the current parameters that is orthogonal to the previous residual direction, ensuring precise enforcement of parameter orthogonality. Then, we integrate $\mathcal{L}_{orth}^i$ into the total unlearning objective:

$$\mathcal{L}^i = \mathcal{L}_u^i \; + \; \lambda\, \mathcal{L}_{orth}^i, \tag{10}$$

where λ balances continual identity unlearning against the orthogonality constraint.

The final optimization is obtained by minimizing $\mathcal{L}^i$, thereby ensuring each new unlearning request preserves all prior unlearning effects. Overall, the *Residual Orthogonal Resonator* significantly reduces identity conflicts in the generative continual identity unlearning while maintaining high visual fidelity.

Phantom Guard Network

To defend against adversarial relearning attacks, we introduce the *Phantom Guard Network* (PGN), which constructs a phantom generator to absorb malicious fine-tuning while preserving the integrity of the unlearned generator. The core idea is to intercept all malicious gradients in a separate phantom copy of the generator, leaving the unlearned generator G_u^i invulnerable. This design provides deceptive feedback with attackers observing apparent identity recovery on the phantom generator, while G_u^i actually retains its unlearning ability. At unlearning request i, let G_u^i be the current unlearned generator with parameters θ_u^i. We instantiate a phantom generator P_u^i by cloning G_u^i:

$$P_u^i(\cdot\,;\phi_u^i) \;\cong\; G_u^i(\cdot\,;\theta_u^i), \quad \phi_u^i \leftarrow \theta_u^i. \tag{11}$$

Then, we freeze the parameters of the unlearned generator G_u^i so that all relearning attacks do not affect G_u^i. Currently, the target image set $\mathcal{T}$ used by $G_s(\cdot;\theta_s)$ is replaced with the relearning attack set A under relearning attacks. We compute the latent codes and their reconstructions under both models:

$$\hat{s}_n = R(P_u^i(E(s_n);\phi_u^i)), \hat{t}_m = R(G_s(E(a_m);\theta_s)), \tag{12}$$

where $\hat{s}_n$ represents the reconstructed image generated for the source image by phantom generator, and $\hat{t}_m$ represents the reconstructed image generated by the target image a_m selected from the relearning attack set A. Then, we define the relearning attack optimization as follows:

$$\phi_u^i \;=\; \arg\min_{\phi} \; \mathcal{L}_{\text{re}}^i(\hat{s}_n, \hat{t}_m, ;\phi), \tag{13}$$

where the parameter θ_u^i of unlearned generator is fixed. This adversarial fine-tuning enables P_u^i to absorb the relearning attack gradients and approximate the attacker's objective without exposing G_u^i. This design provides deceptive feedback, resulting in apparent identity recovery observed by attackers on the phantom generator. After completion of PGN training, we validate the integrity of unlearned generator by reconstructing each original source image s_n through frozen G_u^i:

$$\hat{s}_n = R\big(G_u^i(E(s_n);\theta_u^i)\big). \tag{14}$$

Then, we measure the identity similarity to confirm that the unlearning identities remain irrecoverable between s_n and $\hat{s}_n$. In essence, PGN acts as a sacrificial copy that deflects all relearning attempts, guaranteeing that G_u^i retains its unlearning efficacy. Thus, the *Phantom Guard Network* provides a robust defense that preserves the sanctity of generative continual identity unlearning against relearning attacks. In Figure 1, we provide an overview of ROVER.

Experiments

Experimental Setup

Baseline. Since we propose robust multi-identity unlearning task for the first time, to evaluate the effectiveness of ROVER, we compare it with several continual learning methods, including: Lifelong GAN (Zhai et al. 2019), CO^{2}I (Cha, Lee, and Shin 2021), PIGWM (Zhou et al. 2021), and Selective Amenesia (Heng and Soh 2023). Moreover, we construct the relearning attack settings to test the robustness of the model during generative continual identity unlearning. We use GUIDE (Seo et al. 2024) as the state-of-the-art baseline, which is the first to propose the generative identity unlearning task. To ensure fairness, we evaluate the proposed ROVER and baselines in the same scenario settings.

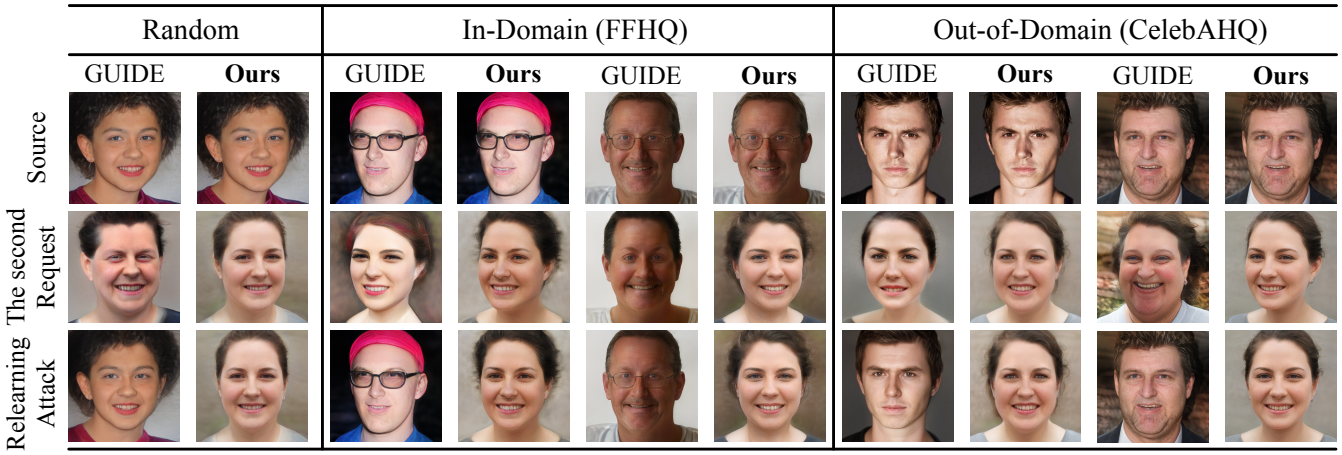


Figure 2: Qualitative results of ROVER and GUIDE in task of robust multi-identity unlearning.

Methods	Random			In-Domain (FFHQ)			Out-of-Domain (CelebAHQ)		
	ID	FID_{pre}	FID_{real}	ID	FID_{pre}	FID_{real}	ID	FID_{pre}	FID_{real}
Lifelong GAN	0.35±0.11	16.78±1.05	9.40±1.19	0.32±0.14	16.19±1.45	8.94±1.20	0.31±0.07	15.72±1.19	8.57±1.03
CO^2I	0.26±0.09	18.16±2.28	9.96±1.20	0.25±0.11	17.23±1.58	9.87±1.08	0.26±0.11	16.31±2.40	9.62±1.25
PIGWM	0.31±0.12	14.56±1.27	8.94±1.08	0.30±0.09	16.87±1.63	9.72±1.40	0.29±0.13	15.63±1.35	9.04±1.31
Selective Amnesia	0.29±0.16	16.31±2.41	9.37±1.35	0.27±0.15	15.20±1.09	8.82±0.96	0.27±0.10	13.48±1.68	7.52±1.42
GUIDE (Continual)	0.14±0.09	12.43±1.34	7.84±1.92	0.12±0.10	13.12±1.47	7.43±1.05	0.14±0.06	8.26±1.21	4.93±0.83
ROVER (Continual)	**0.10±0.02**	**9.85±1.03**	**6.59±1.71**	**0.04±0.03**	**8.19±0.46**	**3.17±0.22**	**0.05±0.03**	**7.67±0.54**	**3.42±0.72**
GUIDE (Relearning)	0.90±0.07	10.28±1.52	6.86±1.04	0.92±0.16	8.56±0.95	3.50±0.63	0.89±0.12	7.85±1.19	3.67±0.70
ROVER (Relearning)	**0.10±0.03**	**9.81±1.27**	**6.62±1.59**	**0.04±0.05**	**8.23±0.68**	**3.14±0.49**	**0.05±0.04**	**7.62±0.85**	**3.45±0.52**

Table 1: Quantitative results of ROVER and baselines in the robust multi-identity unlearning task.

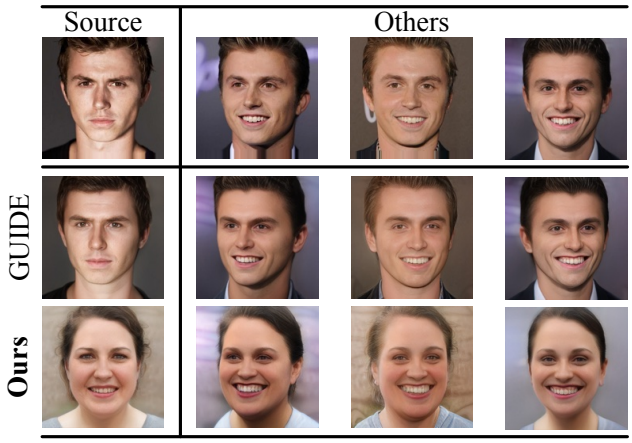


Figure 3: Qualitative results of ROVER and GUIDE.

Implementation Details. We construct ROVER based on the 3D generative adversarial network (Chan et al. 2022) pre-trained on the FFHQ dataset (Karras, Laine, and Aila 2019), which uses GOAE (Yuan et al. 2023) as a GAN inversion network to obtain latent codes from images. The image resolution used in the experiment is 512x512 and the rendering resolution is 128x128. We used the Adam (Kingma and Ba 2015) optimizer for parameter optimization with the learning rate set to 10^{-4}. The boundary threshold ϵ of *Similarity Navigator* is set to 0.5. The orthogonality constraint λ_{orth} in the experiment is set to 0.01.

Datasets and Scenarios. We evaluate ROVER by conducting the task of robust multi-identity unlearning in three scenarios: *Random* sets an unlearning target image from a randomly sampled noise vector; *In-D (In-Domain)* samples a target image from FFHQ dataset (Karras, Laine, and Aila 2019) for model training; *OOD (Out-of-Domain)* samples the unlearning target image from CelebAHQ dataset (Karras et al. 2018). Since there are multiple images with the same identity in CelebAHQ, we also conduct experiments in the *OOD* scenario to test unlearning a single image to evaluate other images with the same identity.

Evaluation Metrics. We evaluate ROVER from the two perspectives of the model's ability to unlearn identities and the overall preservation ability of the generator. Firstly, we use CurricularFace (Huang et al. 2020) to quantitatively measure identity (ID) similarity to assess identity differences generated from the same latent code before and after unlearning. Moreover, we utilize ID_{others} to evaluate the identity erasure ability on images that are not involved in

Methods	The Third Request			The Fourth Request			The Fifth Request		
	ID	FID_{pre}	FID_{real}	ID	FID_{pre}	FID_{real}	ID	FID_{pre}	FID_{real}
GUIDE	0.17±0.04	9.45±1.04	6.03±1.12	0.19±0.06	9.81±1.25	6.89±1.48	0.23±0.09	11.22±1.19	7.57±0.92
ROVER	**0.09±0.02**	**8.02±0.74**	**4.59±1.20**	**0.12±0.07**	**9.68±0.79**	**5.08±0.64**	**0.15±0.03**	**10.73±0.92**	**6.18±0.73**

Table 2: Quantitative results of multiple benign continual unlearning requests on *OOD* scenario using CelebAHQ dataset.

training but contain the same identity as the source image. Secondly, we use two variants of Fréchet Inception Distance (FID) score (Heusel et al. 2017) to evaluate whether our method preserves overall generation performance. Specifically, FID_{pre} evaluates the distribution shift of the images generated by the pre-trained generator and the unlearned generator, while FID_{real} evaluates the distribution shift of the real FFHQ images and the images generated by the unlearned generator. We conducted all experiments 10 times and reported average value, with lower values being better.

Quantitative Experiments

As shown in Table 1, we conduct a comprehensive evaluation of the proposed ROVER in the task of robust multi-identity unlearning. Specifically, we conduct continual unlearning three times in the overall experimental settings and perform a relearning attack during the third round. For the benign continual unlearning requests, we report the average performance of the generated images after the first two rounds of unlearning training. For the relearning attack settings, we report the results after completing all three rounds.

We compare ROVER with several state-of-the-art methods, including both several continual learning methods and the first generative identity unlearning method GUIDE. According to the experimental results, we have the following observations. Firstly, compared to the continual learning methods under benign continual unlearning requests, ROVER achieves significant performance improvements in ID similarity and guarantees high generator preservation ability in three scenarios. Secondly, the SOTA baseline GUIDE performs poorly in both continual unlearning and relearning attack settings, with a particularly severe degradation in the latter, where its ability to unlearn image identities is nearly lost. Overall, our model achieves SOTA performance in benign continual unlearning requests while maintaining the model's performance under relearning attack, which demonstrates the effectiveness and robustness of the proposed ROVER.

To further demonstrate the robustness of ROVER, we represent quantitative results of multiple benign continual unlearning requests on the CelebAHQ dataset, as shown in Table 2. Based on the results, we observe that as the number of unlearning requests increases, the ability to erase identities of previously unlearned images gradually declines, resulting in performance degradation. In contrast, our proposed ROVER achieves robust continual unlearning by mitigating this degradation through residual orthogonality. The results of continual learning methods reported in the appendix.

	ID	ID_{others}	FID_{pre}	FID_{real}
w/o Sim	0.27±0.13	0.51±0.19	10.74±1.52	6.49±1.48
w/o PGN	0.21±0.10	0.34±0.13	9.06±0.67	4.80±0.76
w/o O-Res	0.16±0.06	0.27±0.09	7.87±1.53	3.69±0.90
ROVER	**0.05±0.04**	**0.20±0.12**	**7.62±0.85**	**3.45±0.52**

Table 3: Ablation study of ROVER.

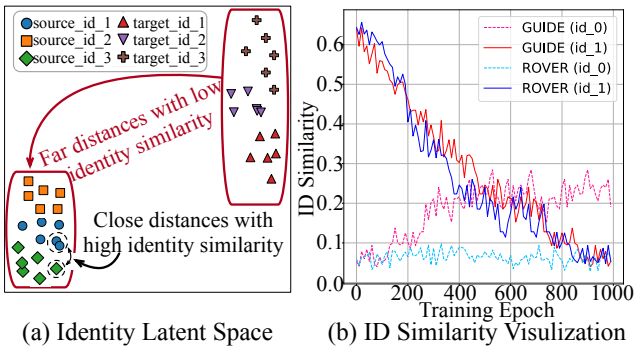
Qualitative Experiments

We conduct a comparative analysis between ROVER and the baseline method GUIDE in the task of robust multi-identity unlearning. Starting from the provided source image, our goal is to erase its identity from the pre-trained generator, as illustrated in Figure 2. We present the unlearned image produced by the model for the first source image after continual unlearning of the second source image, along with the corresponding image generated after the relearning attacks. Notably, ROVER effectively guarantees the ability to erase identities, whether during continual unlearning or under relearning attacks. However, after completing the continual unlearning of the second source image, GUIDE generates distorted output image for the first source image across all evaluated scenarios. In particular, GUIDE nearly regain its ability to reconstruct image identities after the relearning attacks, generating output images that are almost the same as the source images. These results indicate that the proposed ROVER effectively resolves conflict issues during the continual unlearning of multiple images and maintains robust performance under relearning attacks.

To evaluate the thoroughness of identity removal, we use the CelebAHQ dataset to evaluate the effect of unlearning a single image on other images with the same identity. This experiment evaluates ID similarity not only for the unlearned images of the source image but also for other images with the same identity. Figure 3 shows the qualitative results of this experiment. The results indicate that GUIDE recovers not only the identities of the source images but also those of previously unseen images with the same identities after the relearning attacks. However, ROVER successfully retains the ability to unlearn the identities of all images in this experimental setting. This further demonstrates the robustness and effectiveness of the proposed ROVER method in removing identities during continual unlearning against relearning attacks.

Ablation Study

Ablation results. To explore how each component contributes to the performance of ROVER, we conduct a se-



(a) Identity Latent Space (b) ID Similarity Visulization

Figure 4: (a) The distance relationship between the identities of source images and target images in the latent space with t-SNE. (b) Comparison of the visualization of identity unlearning effects between the current image and the previous image during the second round of benign continual unlearning by GUIDE and ROVER.

ϵ	ID	ID_{others}	FID_{pre}	FID_{real}
0.3	0.25±0.16	0.39±0.11	10.49±1.03	7.42±0.93
0.5	**0.05±0.04**	**0.20±0.12**	**7.62±0.85**	**3.45±0.52**
0.7	0.21±0.13	0.29±0.08	11.62±1.27	8.40±0.83
λ_{orth}	ID	ID_{others}	FID_{pre}	FID_{real}
0.001	0.09±0.02	0.27±0.14	12.49±0.14	6.10±0.68
0.01	**0.05±0.04**	**0.20±0.12**	**7.62±0.85**	**3.45±0.52**
0.1	0.11±0.06	0.28±0.19	9.80±0.92	5.98±0.74
0.5	0.16±0.02	0.31±0.11	12.16±0.60	7.45±0.83
1.0	0.21±0.03	0.34±0.12	10.68±0.94	6.42±0.39

Table 4: Parameter analysis.

ries of ablation experiments on *OOD* scenario on the CelebAHQ dataset by excluding each module, as shown in Table 3. We compare ROVER with its variants after three continual unlearning requests, including the final relearning attack: **w/o Sim** indicates that the similarity navigator module is removed, and randomly choose one of PGN and O-Res for model training. **w/o PGN** indicates that the phantom guard network module is removed, and only the residual orthogonal resonator module is used. **w/o O-Res** denotes the removal of the residual orthogonal resonator module, and only the phantom guard network module is used. **ROVER** represents the use of the residual orthogonal resonator module and the phantom guard network module, corresponding to our complete model. From the results, we can observe that the complete model consistently outperforms its single-module counterparts in all evaluation metrics. Besides, randomly selecting training modules compromises the model's ability to erase certain identities or fails to fully prevent relearning attacks. Moreover, the residual orthogonal resonator module makes a greater contribution to our model. The results indicate that resolving identity conflicts during the continual unlearning of multiple images is crucial to achieving effective generative continual identity unlearning.

Necessity analysis. To demonstrate the necessity of ROVER's ID-similarity filtering, we visualize the latent codes of selected source and target images using t-SNE, as shown in Figure 4(a). The source image forms dense clusters with high pairwise ID similarity, indicating that attackers can easily sample from this local neighborhood to replace the target image for training to recover the erasure identity. Under normal circumstances, the target image is located in a distant region of the embedding space with low ID similarity. This clear separation in the latent space highlights the effectiveness and necessity of request filtering based on identity similarity.

Stability analysis. Figure 4(b) present the ID similarity variation between source image and generated image for two identities GUIDE and ROVER during the second unlearning round, which id_0 and id_1 are the unlearned identities in the previous round and current round, respectively. Under ROVER, the current ID similarity gradually decreases, indicating correct forgetting, while the ID similarity of the previous image remains stable at a low level, confirming that early forgetting has not deteriorated. In contrast, GUIDE also reduces the ID similarity of the current round image, but the ID similarity of previous images drifts upward during training, revealing unstable continual unlearning. These results demonstrate that ROVER achieves stable removal of each identity and preserves the integrity of all the previous unlearning steps.

Parameter Analysis

To evaluate the robustness of ROVER under varying hyperparameter configurations, we conduct a parameter sensitivity analysis of ϵ within the range of [0.3, . . ., 0.7] and λ_{orth} within the range of [0.001, . . ., 1.0] on the CelebAHQ dataset. Table 4 presents the experimental results. Based on the observations in the results, we make several conclusions. When the parameters are set too high or too low, the generator's unlearning and preservation capabilities both demonstrate suboptimal performance. ROVER achieves the best trade-off at $\epsilon = 0.5$ and $\lambda_{orth} = 0.01$, effectively unlearning the current round image while preserving the ability to unlearn previous images, which proves its robustness in generative continual identity unlearning.

Conclusion

In this paper, we introduce a novel task called robust multi-identity unlearning. This task requires the complete removal of the multiple identities under continual unlearning requests and defends against relearning attacks. To achieve this, we propose a novel framework, RObust generatiVE continual identity unlearning against Relearning attacks (ROVER), which resolves identity conflicts arising from the benign continual unlearning and prevents malicious relearning attacks with filtering unlearning requests. The experimental results show that ROVER can effectively accomplish the task of robust multi-identity unlearning, which verifies the effectiveness and robustness of the proposed framework.

Acknowledgements

This research is funded by National Natural Science Foundation of China (No. 62406347). Y. Chen was supported by the Hong Kong Research Grants Council, Early Career Scheme Fund [Grant 26508924].

References

Cha, H.; Lee, J.; and Shin, J. 2021. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9516–9525.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Chen, J.; Li, J.; Song, C.; Li, B.; Chen, Q.; Gao, H.; Wang, W. H.; Xu, Z.; and Shi, X. 2024. Discriminative forests improve generative diversity for generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11338–11345.

Chu, T.; Xing, W.; Chen, J.; Wang, Z.; Sun, J.; Zhao, L.; Chen, H.; and Lin, H. 2024. Attack deterministic conditional image generative models for diverse and controllable generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1362–1370.

Cong, Y.; Zhao, M.; Li, J.; Wang, S.; and Carin, L. 2020. Gan memory with no forgetting. *Advances in neural information processing systems*, 33: 16481–16494.

Dai, Z.; and Gifford, D. K. 2023. Training data attribution for diffusion models. *arXiv preprint arXiv:2306.02174*.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2024. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.

Gao, H.; Pang, T.; Du, C.; Hu, T.; Deng, Z.; and Lin, M. 2024. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*.

He, S.; Wang, R.; Liu, T.; Yi, C.; Jin, X.; Liu, R.; and Zhou, W. 2022. Type-I generative adversarial attack. *IEEE Transactions on Dependable and Secure Computing*, 20(3): 2593–2606.

Heng, A.; and Soh, H. 2023. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36: 17170–17194.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hu, H.; and Pang, J. 2021. Stealing machine learning models: Attacks and countermeasures for generative adversarial networks. In *Proceedings of the 37th Annual Computer Security Applications Conference*, 1–16.

Huang, C.-P.; Chang, K.-P.; Tsai, C.-T.; Lai, Y.-H.; Yang, F.-E.; and Wang, Y.-C. F. 2024a. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, 360–376.

Huang, L.; An, Z.; Zeng, Y.; Xu, Y.; et al. 2024b. Kfc: Knowledge reconstruction and feedback consolidation enable efficient and effective continual generative learning. In *The Second Tiny Papers Track at ICLR 2024*.

Huang, T.; Wang, Y.; Li, Q.; He, C.; and Gao, J. 2025. Can LLMs Find Fraudsters? Multi-level LLM Enhanced Graph Fraud Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1530–1538.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5901–5910.

Jones, E.; Dragan, A.; Raghunathan, A.; and Steinhardt, J. 2023. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning*, 15307–15329. PMLR.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of gans for improved quality, stability and variation. In *International Conference on Learning Representations*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22691–22702.

Lee, S.; and Min, M. 2025. CG-TGAN: Conditional Generative Adversarial Networks with Graph Neural Networks for Tabular Data Synthesizing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18145–18153.

Li, Y.; Bradshaw, J.; and Sharma, Y. 2019. Are generative classifiers more robust to adversarial attacks? In *International Conference on Machine Learning*, 3804–3814. PMLR.

Li, Z.; Zhao, X.; Wu, D.-D.; Cui, J.; and Shen, Z. 2025. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *arXiv preprint arXiv:2503.10635*.

Liu, S.; and Tan, Y. 2024. Unlearning Concepts from Text-to-Video Diffusion Models. *arXiv preprint arXiv:2407.14209*.

Moon, S.; Cho, S.; and Kim, D. 2024. Feature unlearning for pre-trained gans and vaes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21420–21428.

Panda, S.; and Prathosh, A. 2024. FAST: Feature Aware Similarity Thresholding for weak unlearning in black-box generative models. *IEEE Transactions on Artificial Intelligence*.

Qiang Chen, A. H. X. L. S. J. S. Y. C. X. Y. C. X. S., Zhongze Wu. 2025. Graph Unlearning Meets Influence-aware Negative Preference Optimization. In *ACM MM*.

Saxena, D.; Cao, J.; Xu, J.; and Kulshrestha, T. 2024. Rg-gan: Dynamic regenerative pruning for data-efficient generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4704–4712.

Seo, J.; Lee, S.-H.; Lee, T.-Y.; Moon, S.; and Park, G.-M. 2024. Generative unlearning for any identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9151–9161.

Sun, H.; Zhu, T.; Chang, W.; and Zhou, W. 2025. Generative adversarial networks unlearning. *IEEE Transactions on Dependable and Secure Computing*.

Tiwary, P.; Guha, A.; Panda, S.; et al. 2023. Adapt then unlearn: Exploiting parameter space semantics for unlearning in generative adversarial networks. *arXiv preprint arXiv:2309.14054*.

Wang, X.; Yang, G.; Ye, T.; and Liu, Y. 2025. DehazeRetinexGAN: Real-World Image Dehazing via Retinex-based Generative Adversarial Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7997–8005.

Wu, C.; Luo, W.; Zhou, N.; Xu, P.; and Zhu, T. 2021. Genetic algorithm with multiple fitness functions for generating adversarial examples. In *2021 IEEE Congress on evolutionary computation (CEC)*, 1792–1799. IEEE.

Wu, J.; Le, T.; Hayat, M.; and Harandi, M. 2024. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*.

Wu, Y.; Zhou, S.; Yang, M.; Wang, L.; Chang, H.; Zhu, W.; Hu, X.; Zhou, X.; and Yang, X. 2025. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8496–8504.

Xie, Y.; Li, Z.; Shi, C.; Liu, J.; Chen, Y.; and Yuan, B. 2021. Enabling fast and universal audio adversarial attack using generative model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14129–14137.

Xu, Y.; Liang, J.; Jia, G.; Yang, Z.; Zhang, Y.; and He, R. 2023. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22658–22668.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22412–22423.

Yang, C.; Wu, Q.; Li, H.; and Chen, Y. 2017. Generative poisoning attack method against neural networks. *arXiv preprint arXiv:1703.01340*.

Ye, F.; and Bors, A. G. 2022a. Continual variational autoencoder learning via online cooperative memorization. In *European Conference on Computer Vision*, 531–549. Springer.

Ye, F.; and Bors, A. G. 2022b. Lifelong generative modelling using dynamic expansion graph model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8857–8865.

Yong, Z.-X.; Menghini, C.; and Bach, S. H. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Yuan, Z.; Zhu, Y.; Li, Y.; Liu, H.; and Yuan, C. 2023. 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*, 2437–2447.

Zhai, M.; Chen, L.; and Mori, G. 2021. Hyper-lifelonggan: Scalable lifelong learning for image conditioned generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2246–2255.

Zhai, M.; Chen, L.; Tung, F.; He, J.; Nawhal, M.; and Mori, G. 2019. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2759–2768.

Zhang, G.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2024a. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1755–1764.

Zhang, Y.; Chen, X.; Jia, J.; Zhang, Y.; Fan, C.; Liu, J.; Hong, M.; Ding, K.; and Liu, S. 2024b. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in Neural Information Processing Systems*, 37: 36748–36776.

Zhang, Y.; Guo, K.; and Zhou, X. 2024. Causally aware generative adversarial networks for light pollution control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22529–22537.

Zhang, Y.; Jia, J.; Chen, X.; Chen, A.; Zhang, Y.; Liu, J.; Ding, K.; and Liu, S. 2024c. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, 385–403.

Zhang, Y.; Jia, R.; Pei, H.; Wang, W.; Li, B.; and Song, D. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 253–261.

Zhou, J.; Wu, Z.; Xue, Y.; Li, M.; and Zhou, D. 2022. Network unknown-threat detection based on a generative adversarial network and evolutionary algorithm. *International Journal of Intelligent Systems*, 37(7): 4307–4328.

Zhou, M.; Xiao, J.; Chang, Y.; Fu, X.; Liu, A.; Pan, J.; and Zha, Z.-J. 2021. Image de-raining via continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4907–4916.