Discriminator-Guided Unlearning: A Framework for Selective Forgetting in Conditional GANs

Byeongcheon Lee[1], Sangmin Kim[1], Sungwoo Park[2], Seungmin Rho[2] and Mi Young Lee[2,*]

[1]Department of Security Convergence, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Department of Industrial Security, Chung-Ang University, Seoul 06974, Republic of Korea

Abstract

The advancement of generative artificial intelligence (AI) has made the machine unlearning essential for resolving data privacy and copyright issues. While retraining models from scratch is effective, it is computationally expensive, and fine-tuning, a common alternative, suffers from catastrophic forgetting. To overcome this problem, this study proposes a two-step framework for auxiliary classifier-based generative adversarial network called "Discriminator-Guided Unlearning." Instead of directly targeting the generator, we intentionally weaken the ability of the discriminator to recognize specific classes. Feedback from this weakened discriminator guides the generator to avoid generating images of that class. Experiments suggest that our framework achieves effective forgetting performance comparable to retrained models while maintaining the quality of the remaining classes, effectively mitigating "catastrophic forgetting." Our approach provides a promising direction for building trustworthy AI.

Keywords

Machine Unlearning, Image Generation Model, Generative Adversarial Networks, Selective Forgetting, Data Privacy, AI Safety

1. Introduction

Generative artificial intelligence (AI) techniques, such as generative adversarial networks (GANs) [1], have achieved remarkable success in various computer vision domains [2, 3]. These generative AI techniques require massive datasets for training to produce high-quality outputs. However, this poses serious ethical and legal issues, as the models can indiscriminately learn and reproduce copyrighted images, sensitive personal information, and biased data [4]. These issues, in conjunction with the need for regulations such as the "right to be forgotten" in the General Data Protection Regulation (GDPR), are further increasing demand for "machine unlearning," the technique of removing the influence of specific data from trained models [5].

In machine unlearning, the simplest approach is to remove the data to be forgotten and then retrain the model from scratch using the remaining dataset. While this approach has the advantage of completely removing desired information, the enormous time and computational resources required for training make it impractical [6]. A more practical alternative approach is to fine-tune a trained model with the remaining dataset, but this approach may lead to "catastrophic forgetting" problem [7]. This problem not only removes target information but also loses knowledge about other data that the model needs to retain, significantly degrading overall performance.

Therefore, we propose a practical and effective selective unlearning framework that avoids both the excessive cost of retraining and the performance degradation of fine-tuning. The proposed framework is composed of a novel two-step method of “discriminator-guided unlearning," which is specialized for auxiliary classifier-based GAN (ACGAN) [8]. In the first step of the proposed framework, the ability of a discriminator to recognize a particular class is intentionally weakened, and in the second step, this ‘confused’ discriminator is utilized to guide the generator to stop generating images of that class. The main contributions of this study are as follows:

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.
qudcjs0208@cau.ac.kr (B. Lee); kimddol98@cau.ac.kr (S. Kim); psw5574@cau.ac.kr (S. Park); smrho@cau.ac.kr (S. Rho); miylee@cau.ac.kr (M. Y. Lee)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-16

CEUR-WS.org/Vol-4132/short25.pdf

1. Unlike existing unlearning methods that focus on directly modifying the generator, we propose a novel unlearning mechanism that induces confusion in the discriminator to prevent recognition of specific classes and then uses its feedback to guide the generator away from producing images of the target classes.
2. Through extensive experiments on representative image datasets including MNIST, FashionMNIST, SVHN, and CIFAR-10, we indicate that the proposed framework can mitigate the catastrophic forgetting problem by suppressing the generation of classes to be forgotten while maintaining high-quality image generation for classes to be retained.

The remainder of this paper is organized as follows. In Section 2, we introduce the theoretical background of GANs and machine unlearning methods. In Section 3, we describe the details of the proposed framework and then present experimental results in Section 4. Finally, in Section 5, we summarize our study and present the conclusions.

2. Theoretical Background

This section briefly reviews GANs and machine unlearning methods, which are the core technologies of our research, and explains why we adopted ACGAN and a discriminator-based unlearning strategy.

2.1. Generative Adversarial Networks (GANs)

GANs, first proposed by Goodfellow [1], utilize a competitive process of a generator, which produces a synthetic image from noise, and a discriminator, which distinguishes between real and synthetic images. While GANs have significantly advanced the field of generative modeling [2], the primary structure of a GAN is limited by the fact that it uses only random noise as input, making it difficult to control the data generation process precisely. To overcome this limitation, conditional GANs (cGANs) [9] have been proposed. cGANs leverage additional condition information, such as class labels, within both the generator and discriminator to control the image generation process in a desired direction.

In this study, we adopted the ACGAN [8] as the primary model for the proposed framework to further maximize the advantages of cGAN. ACGAN is a structure that adds an auxiliary classifier to the cGAN, which allows the discriminator to determine the authenticity of an image while also classifying the generated image. As a result, both the quality of the generated image and the learning stability of the GAN are greatly improved. Furthermore, since the core of our research is "discriminator-driven unlearning," which aims to weaken the ability of a discriminator to distinguish between specific classes, the ACGAN structure, with its explicit classification function within the discriminator, is well-suited to our research objective.

2.2. Machine Unlearning

Machine unlearning research explores how to efficiently remove the influence of specific data from trained models. As comprehensively surveyed by [10], research has shifted to more efficient "approximate unlearning" methodologies to overcome the practical limitations of retraining.

Most of these methodologies focus on directly modifying the generator to induce forgetting. These include improving the fine-tuning process via techniques like label reversal [11], using gradient-descent-based methodologies [12], directly altering model parameters or features [13], and utilizing knowledge distillation [14]. This line of inquiry continues to evolve, with recent work by [15], for instance, focusing on unlearning specific identities from the latent space of a GAN. While many methods target the generator, other components of the machine learning pipeline have also been leveraged. A different line of work by [16], for example, explores using an auxiliary discriminator trained for membership inference as a guiding signal for unlearning in classification models.

However, most of these methodologies risk degrading the performance of other classes when removing information from the target class. Therefore, we propose a novel approach that allows for effective

unlearning while preserving the quality of generation for other classes by initially focusing on the discriminator.

3. Proposed Framework

In this paper, we propose a novel two-step selective unlearning framework for ACGAN models to effectively remove certain forgetting classes c_f. The overall flow of the proposed framework is summarized in Figure 1. In the first step, the goal is to "confuse" the discriminator (D_{forget}) so that it cannot clearly distinguish the forgetting class (c_f). In the second step, we utilize the feedback from this confused discriminator to guide the pre-trained original generator ($G_{original}$) to no longer generate images of class c_f.

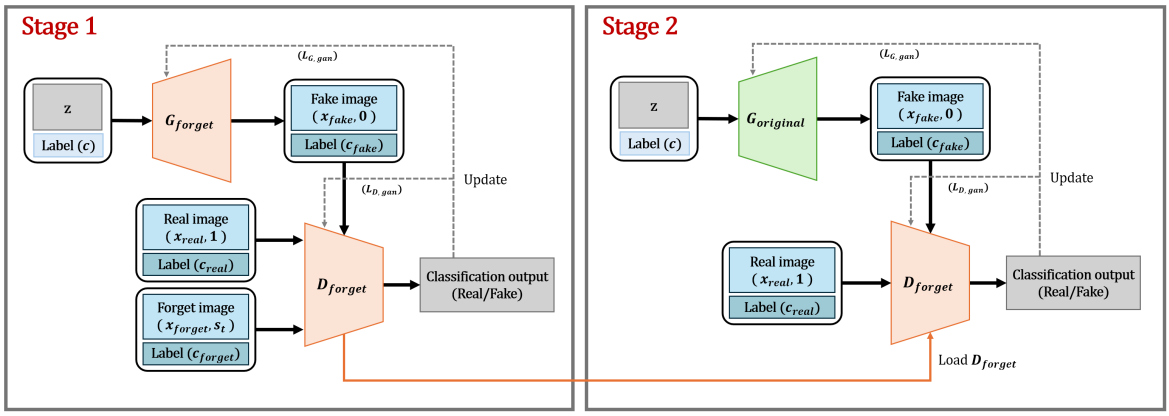

Figure 1: Overview of the Proposed Two-Step Unlearning Framework.

3.1. Step 1: Discriminator Soft Forgetting

The first step of unlearning is the "soft forget" process described in the left panel of Figure 1. In this step, a new generator (G_{forget}) and a new discriminator (D_{forget}) are trained together from scratch. The main goal of this phase is to selectively weaken the ability of the discriminator to determine the forgetting class c_f. To do this, the source discrimination loss is modified when a true image (x_f) that belongs to the forgetting class c_f is input to the discriminator.

In the ACGAN in this experiment with hinge loss, discriminator D is trained to output $D_s(x_r) \geq 1$ for the real image x_r and $D_s(x_{fake}) \leq -1$ for the fake image x_{fake}. In the proposed framework, we introduce a hyperparameter, a 'soft target' $s_t \in [0, 1]$, to confuse the discriminator for real images x_f of the forgetting class c_f. The loss function is a weighted sum that treats the image as 'real' with probability s_t and 'fake' with probability $(1 - s_t)$.

Let x_r denote the true image of class c_r, x_f denote the true image of the forgetting class c_f, z denote the latent vector, and D_s denote the source output of the discriminator. In the third term (the unlearning term), if $s_t = 1$, the loss is equal to the hinge loss. If $s_t = 0$, the discriminator is trained to classify real images of class c_f as fake. Through this process, we ultimately obtain a discriminator (D_{forget}) whose judgment on class c_f is intentionally weakened.

Mathematically, the total GAN loss for the discriminator, $L_{D,gan}$, is expressed as follows:

$$\begin{aligned} L_{D,gan} = & \underbrace{\mathbb{E}_{x_r, c \neq c_f}[\max(0, 1 - D_S(x_r))]}_{\text{Real (Non-Forget)}} + \underbrace{\mathbb{E}_{z,c'}[\max(0, 1 + D_S(G(z, c')))]}_{\text{Fake (All Classes)}} \\ & + \underbrace{\mathbb{E}_{x_f}[s_t \cdot \max(0, 1 - D_S(x_f)) + (1 - s_t) \cdot \max(0, 1 + D_S(x_f))]}_{\text{Real (Forget Class)}} \end{aligned} \tag{1}$$

3.2. Step 2: Generator and Discriminator Final Fine-tuning

The second step corresponds to the right panel of Figure 1. The goal of this step is to complete the final training using the 'confused' discriminator (D_{forget}) obtained in step 1, and to ensure that the pre-trained original generator ($G_{original}$) no longer produces images of the forgetting class c_f.

In this step, the loss function of the discriminator remains the same as in step 1, which means that the discriminator continues to be trained in a confused state for class c_f. The key change is in the loss function of the generator. When the generator attempts to generate an image of the forgetting class c_f, it is penalized in the opposite direction to the standard GAN loss. The GAN loss for the generator $L_{G,gan}$, is split into two parts.

$$L_{G,gan} = \underbrace{\mathbb{E}_{z,c\neq c_f}[-D_S(G(z,c))]}_{\text{Standard loss for normal classes}} + \underbrace{\mathbb{E}_z[D_S(G(z,c_f))]}_{\text{Penalty for forget class}} \tag{2}$$

The first term is the standard hinge loss, which trains the output D_s of the discriminator to be closer to +1 (to look real) when the generator produces an image of a class other than c_f. The second term, on the other hand, teaches the generator to minimize the output D_s of the discriminator for class c_f, i.e., to get closer to -1 (look more fake). Since D_{forget} already tends to give a low score for class c_f, the generator will naturally avoid generating images of this class in the process of minimizing this penalty term. The total loss of the generator is $L_G = L_{G,gan} + w_{aux} \cdot L_{G,aux}$.

4. Experimental Results

This section details the experimental setup and experimental results. Especially, we evaluated the proposed framework by four aspects: qualitative, quantitative, runtime (efficiency), and attack-based robustness.

4.1. Experimental Setup

Our experiments were conducted on the representative image generation benchmark datasets MNIST, FashionMNIST, SVHN, and CIFAR-10. All experiments were performed on a system equipped with two NVIDIA RTX 5000 Ada Generation GPUs. To ensure consistency and reliability, we utilized the same ACGAN architecture in all experiments, which combines the structural advantages of ResNet [17] with self-attention [18]. The main hyperparameters used in the experiments are summarized in Table 1.

Table 1
Key hyperparameters for training and optimization.

Category	Value
Optimization	
Optimizer	Adam
Learning rate (Generator)	2×10^{-5}
Learning rate (Discriminator)	4×10^{-5}
Adam Betas (β_1, β_2)	(0.5, 0.999)
Batch size	512
Soft target (s_t)	0.1
Training Epoch	
Original / Retrain	200
Step 1 (Soft Forgetting)	200
Step 2 / Finetune	50

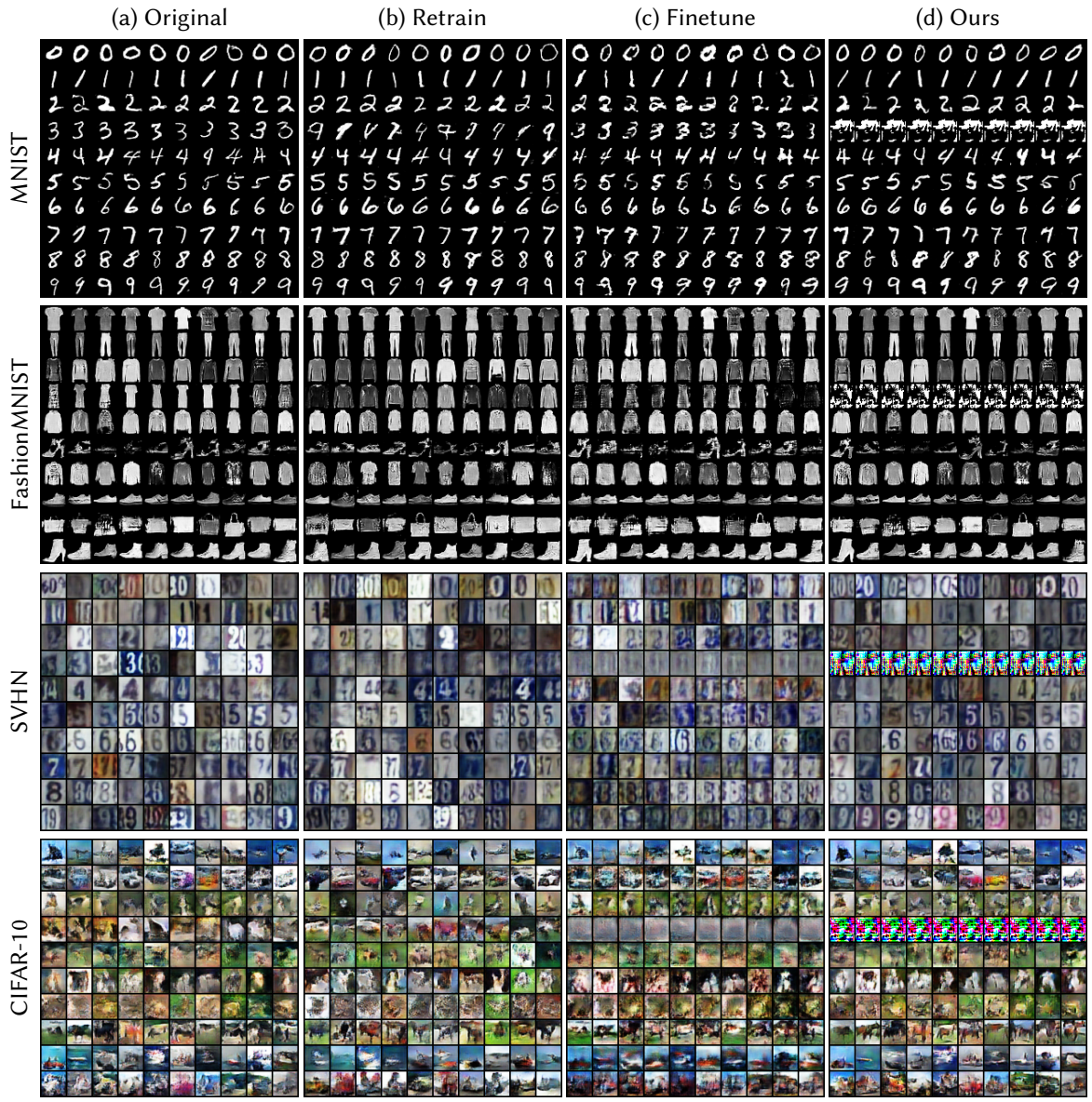

Figure 2: Qualitative comparison of unlearning methods (columns).

4.2. Qualitative Results

To analyze visually the effectiveness of the proposed discriminator-guided unlearning framework, we compare images generated using a fixed noise vector for each model. Figure 2 shows these qualitative results. Each row corresponds to a different dataset (MNIST, SVHN, Fashion-MNIST, CIFAR-10), and each column represents the output of a different model: (a) Original, (b) Retrain, (c) Finetune, (d) Our proposed model. In all experiments, specific classes are targeted for forgetting (‘3’ in MNIST, ‘3’ in SVHN, ‘coat’ in Fashion-MNIST, ‘cat’ in CIFAR-10).

The results consistently suggest the effectiveness of the proposed framework. As shown in columns (a) and (b), the original model successfully generates all classes, and the retrained model suppresses the generation of forgotten classes while maintaining high quality for the remaining classes. In contrast, the fine-tuned model (c) exhibited a "catastrophic forgetting" phenomenon. This phenomenon is most evident in the complex CIFAR-10 dataset, where a degradation in image generation quality for the remaining class is observed.

The proposed framework (d) successfully unlearns the forgetting target class in all datasets, producing images of that class with unrecognizable noise patterns. Moreover, the image quality of the remaining

Table 2
Overall performance comparison of unlearning methods across datasets.

Dataset	Model Type	Forgetting (Target Only)			Retention (Exclude Target)		
		FID(↑)	KID(x100)(↑)	IS(↓)	FID(↓)	KID(x100)(↓)	IS(↑)
MNIST	Original	63.76	6.19	1.55	6.46	0.28	2.16
	Retrain	39.81	3.4	2.06	6.99	0.31	2.19
	Finetune	85.3	9.31	1.48	26.46	2.28	2.17
	Ours	254.35	32.74	1.07	5.27	0.28	2.13
FashionMNIST	Original	92.68	7.49	2.86	21.48	1.02	4.22
	Retrain	97.2	7.11	2.26	24.35	1.41	4.26
	Finetune	131.3	11.95	2.82	40.25	2.59	4.13
	Ours	417.23	52.88	1.04	19.63	0.84	4.27
SVHN	Original	55.79	4.78	2.72	50.21	4.14	2.82
	Retrain	152.92	14.83	1.69	66.37	6.06	2.63
	Finetune	154.43	14.8	1.59	82.52	6.73	2.52
	Ours	514.34	73.51	1.01	54.74	4.85	2.54
CIFAR-10	Original	123.1	10.7	2.95	83.36	6.17	4.26
	Retrain	126.16	13.16	2.91	78.93	5.81	4.49
	Finetune	169.17	13.56	1.95	110.59	9.11	3.35
	Ours	420.59	50.47	1.02	82.36	6.46	4.14

nine classes is preserved at almost the same level as the retraining model (b). This consistency is maintained regardless of the complexity of the data, from simple black and white numbers in MNIST to complex color objects in CIFAR-10.

In summary, the qualitative results demonstrate that the proposed framework selectively and effectively removes the target class without "catastrophic forgetting."

4.3. Quantitative Results

Table 2 presents quantitative results comparing our framework with the baseline model. To provide a comprehensive evaluation, we utilize three standard metrics: Fréchet Inception Distance (FID), Kernel Inception Distance (KID), and Inception Score (IS). FID is a distance calculated by approximating the feature distributions of the real and generated images as Gaussians in the Inception network space, where lower values indicate greater similarity between the two distributions. KID is a metric that calculates the squared Maximum Mean Discrepancy (MMD^2) based on a polynomial kernel for the Inception features of real and generated images, where lower values indicate greater similarity between the two distributions. IS is an indicator that simultaneously measures the quality and diversity of generated images, where higher scores are preferable. Their formulations are as follows:

$$\text{FID}(P_r, P_g) = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{3}$$

$$\text{MMD}^2(P_r, P_g) = \mathbb{E}_{x,x'\sim p_r}[K(x, x')] + \mathbb{E}_{x,x'\sim p_g}[K(x, x')] - 2\mathbb{E}_{x\sim p_r, x'\sim p_g}[K(x, x')] \tag{4}$$

$$\text{IS}(P_g) = \exp(\mathbb{E}_{x\sim p_g}[\text{KL}(p(y|x)||p(y))]) \tag{5}$$

The interpretation of these metrics is nuanced depending on the evaluation scenario. In the retention scenario, the standard interpretation holds, as the goal is to preserve high-quality generation for the remaining classes. Conversely, in the forgetting scenario, the objective is to fail at generating the target class; therefore, a higher FID/KID score is interpreted as more effective unlearning, as it signifies that the generated outputs have successfully collapsed into a noise distribution far from the real data. The results clearly show the superiority of our method in achieving a balance between effective forgetting and knowledge preservation on all test datasets.

In the forgetting scenario, which measures how well the target class is removed, our method achieves a much more complete unlearning effect than all baseline models. The forgetting FID/KID scores are significantly higher, indicating that the images generated for the target class have successfully collapsed into unrecognizable noise. For example, on FashionMNIST, the forgetting FID (417.23) is significantly higher than the fine-tuning (131.30). More importantly, this strong forgetting ability does not lead to "catastrophic forgetting." In the retention scenario, our framework consistently and significantly outperforms the fine-tuning model. For MNIST and FashionMNIST, retention performance is better than the retrained baseline model, suggesting a positive generalization effect from our unlearning process. Even on the more complex CIFAR-10 dataset, performance remains competitive with the retrained model.

4.4. Execution Time

In this section, to demonstrate efficiency, we compare the execution times of retrain, finetune, and our proposed framework. As shown in Table 3, the execution phase of our framework is significantly faster than full retraining, offering a crucial advantage for time-sensitive removal requests. More importantly, unlike fine-tuning approach that offer comparable speed but suffer from catastrophic forgetting phenomenon, our framework maintains high image quality comparable to much slower retraining models.

Table 3
Comparison of Unlearning Execution Times. (seconds)

Dataset	Retrain	Finetune	Ours
MNIST	10379.9	1060.8	770.3
FashionMNIST	10432.3	1044.4	761.5
SVHN	15443.7	1530.4	1675.1
CIFAR-10	12176.6	1219.3	1297.6

4.5. Attack-Based Evaluation

This section presents qualitative evidence from a model inversion attack and quantitative proof from a membership inference attack (MIA). This section aims to more directly and rigorously quantify the absence of specific learned information. A model inversion attack visually probes the conceptual knowledge of the model. In contrast, MIA aims to determine whether a specific data point was part of the original training set of the model by exploiting subtle differences in the output of the model on training data versus unseen data. Therefore, for an unlearned class, an effective unlearning method should obscure these behavioral differences, thereby reducing the attack accuracy.

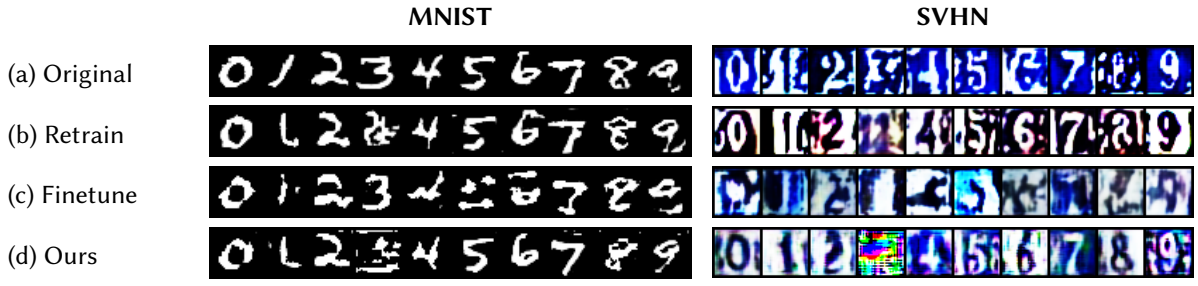

Figure 3: Model Inversion Attack results on MNIST and SVHN. The unlearning target was class '3'.

The qualitative results in Figure 3 are unequivocal. The reconstructed image for the forgotten class '3' in our model (d) is an unrecognizable artifact, mirroring the 'Retrain' gold standard (b). This contrasts with the Finetune model (c), which fails to fully forget the target class and exhibits severe catastrophic

Table 4
Membership Inference Attack Accuracy.

Model Type	MNIST		FashionMNIST		SVHN		CIFAR-10	
	Forget	Retain	Forget	Retain	Forget	Retain	Forget	Retain
Original	0.804	0.784	0.762	0.766	0.693	0.639	0.744	0.737
Retrain	0.766	0.786	0.762	0.76	0.636	0.627	0.729	0.733
Finetune	0.762	0.757	0.758	0.752	0.659	0.642	0.725	0.723
Ours	0.753	0.798	0.757	0.762	0.608	0.657	0.68	0.774

forgetting on retained digits in the SVHN dataset. Our method, however, preserves the quality of all other digits, providing clear visual proof of selective and robust unlearning.

The quantitative MIA results in Table 4 further solidify these findings. Our framework consistently achieves the lowest attack accuracy on the Forget Set—even surpassing the Retrain baseline on complex datasets—while maintaining a high accuracy on the Retain Set. This confirms that our strong forgetting performance does not induce catastrophic forgetting.

5. Conclusions

In this study, we propose a selective unlearning framework based on a "discriminator-guided" method to effectively remove specific classes from generative models. The framework introduces a two-step structure. First, it selectively weakens the ability of the discriminator to recognize the target class and then uses this altered feedback to guide the unlearning process of the generator.

The experimental results validated across four benchmark datasets show that this indirect approach presents a promising solution to the unlearning problem. The framework achieves strong forgetting performance comparable to a retrained model, effectively suppressing the generation of target class images. It successfully mitigates the “catastrophic forgetting" problem associated with simple fine-tuning, maintaining stable image generation quality for retained classes. Furthermore, the effectiveness of the framework was additionally confirmed through attack-based evaluations, including model inversion and membership inference attacks. These findings indicate that discriminator-guided unlearning is a practical and effective direction for removing specific data from generative models, demonstrating a superior balance between performance, efficiency, and verifiability.

However, a practical limitation is that its time efficiency relies on the first step being prepared in advance for anticipated unlearning requests. Future work will focus on applying this framework to other types of generative models and validating its effectiveness on high-resolution image datasets.

Acknowledgments

This research was supported by the "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government. (2025-RISE-01-024-04) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00518960, 50%).

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI ChatGPT-4 and DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[2] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[4] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al., Extracting training data from large language models, in: 30th USENIX security symposium (USENIX Security 21), 2021, pp. 2633–2650.

[5] Q.-V. Dang, Right to be forgotten in the age of machine learning, in: International Conference on Advances in Digital Science, Springer, 2021, pp. 403–411.

[6] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, in: 2021 IEEE symposium on security and privacy (SP), IEEE, 2021, pp. 141–159.

[7] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of learning and motivation, volume 24, Elsevier, 1989, pp. 109–165.

[8] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: International conference on machine learning, PMLR, 2017, pp. 2642–2651.

[9] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[10] A. Huang, Z. Cai, Z. Xiong, A survey of machine unlearning in generative ai models: Methods, applications, security, and challenges, IEEE Internet of Things Journal (2025).

[11] L. Li, P.-g. Ye, Z. Li, Z. Yang, Z. Zhang, Finetune and label reversal: Privacy-preserving unlearning strategies for gan models in cloud computing, Computer Standards & Interfaces 93 (2025) 103976.

[12] A. Golatkar, A. Achille, S. 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, 2020, pp. 9304–9312.

[13] S. Moon, S. Cho, D. Kim, Feature unlearning for pre-trained gans and vaes, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 21420–21428.

[14] H. Kim, S. Lee, S. S. Woo, Layer attack unlearning: Fast and accurate machine unlearning via layer level attack and knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 21241–21248.

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

[16] A. Zhavoronkin, M. Pautov, N. Kalmykov, E. Sevriugov, D. Kovalev, O. Y. Rogov, I. V. Oseledets, Ungan: machine unlearning strategies through membership inference, 540 (2024) 46–60.

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[18] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International conference on machine learning, PMLR, 2019, pp. 7354–7363.