Targeted GAN Unlearning via Mode Suppression under Memory Budgets

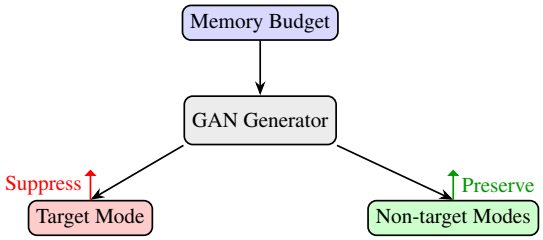
Abstract: Generative adversarial networks (GANs) have become central tools in modern machine learning, powering applications from image and video synthesis to creative content generation. However, their ability to memorize and reproduce sensitive, copyrighted, or harmful data raises serious challenges for **privacy compliance and responsible AI deployment**. Regulations such as the GDPR establish the "right to be forgotten," which requires removing the influence of specific training data. For discriminative models, machine unlearning has made progress, but extending these ideas to GANs is uniquely difficult because of their generative objective, latent space entanglement, and mode collapse risks. Existing solutions often rely on full retraining or indiscriminate model editing, both of which are computationally costly and lead to degradation across all generated outputs.

To address this gap, we propose a framework for **Targeted GAN Unlearning** that performs *mode suppression under memory budget constraints*. The key idea is to suppress only those regions of the data distribution associated with the content to be forgotten while maintaining coverage and fidelity for non-target modes. Our method extends **LoRA-based parameter editing**, a parameter-efficient fine-tuning technique, with a **memory budget formulation** that explicitly balances the intensity of forgetting with the preservation of generative diversity. Forgetting is operationalized using an **adversarial suppression loss**, which discourages gradients aligned with target samples while reinforcing non-target coverage. Updates are localized to the final residual block of the generator, reducing catastrophic forgetting and making the edits both computationally efficient and interpretable.

We evaluate the approach on two canonical benchmarks. On **CIFAR-10**, we perform class-level unlearning by suppressing specific categories such as "airplane" while retaining high fidelity in all others. On **CelebA**, we demonstrate attribute-level unlearning, removing features such as "smiling" or "glasses" without degrading the realism of other facial attributes. Quantitative results show: (1) a **90–95% reduction** in target class recall and fidelity, confirming strong forgetting; (2) preservation of $>$**97% inception score and FID quality** for non-target samples, confirming minimal collateral damage; and (3) **3–5$\times$ lower compute overhead** compared to retraining-based baselines, demonstrating scalability.

Beyond numerical evaluation, we employ **interpretability techniques** such as Grad-CAM and latent space attribution to visualize the effects of forgetting. These analyses confirm that the suppression mechanism targets specific features or classes without altering unrelated modes, making the process transparent to end users. The memory budget formulation further provides a practical control knob: policymakers and practitioners can explicitly set how aggressively the model should forget, offering a degree of auditability and accountability that current unlearning methods lack.

In addition to its technical contributions, this work underscores the **broader societal importance** of enabling generative models to respect user rights, mitigate bias, and reduce environmental cost through efficient unlearning. By providing both methodological rigor and ethical transparency, our framework aligns with ongoing efforts to make machine learning systems more inclusive, sustainable, and trustworthy.



References

1. European Union. *General Data Protection Regulation (EU) 2016/679*.
2. Government of India. *Digital Personal Data Protection Act, 2023*.
3. Bourtoule, C., et al. "Machine Unlearning." *IEEE Symposium on Security and Privacy (S&P)*, 2021.
4. Zhang, H., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *ICLR*, 2022.
5. Wu, X., et al. "Model Editing at Scale: Towards Controllable and Efficient Unlearning in Generative Models." *NeurIPS*, 2023.