Reproduction of RLB-MI: Reinforcement Learning-based Black-Box Model Inversion Attack

CSEG516 Reinforcement Learning Term Project

Reproduction Study | Reinforcement Learning | Privacy Attack | GAN + ArcFace

Team: Daewon Kim, Seobin Choi

Sogang University, Fall 2025

Abstract: *This project reproduces the RLB-MI (Reinforcement Learning-based Black-Box Model Inversion) attack framework from CVPR 2023. We implement Soft Actor-Critic (SAC) to search the latent space of a pre-trained GAN, targeting ArcFace-based face classifiers trained on CelebA and FaceScrub datasets. Additionally, we investigate the effect of* ***inference-time logit temperature scaling*** *(varying the ArcFace scale factor s) on attack performance, demonstrating that calibrated confidence scores lead to more informative reward signals for the RL agent.*

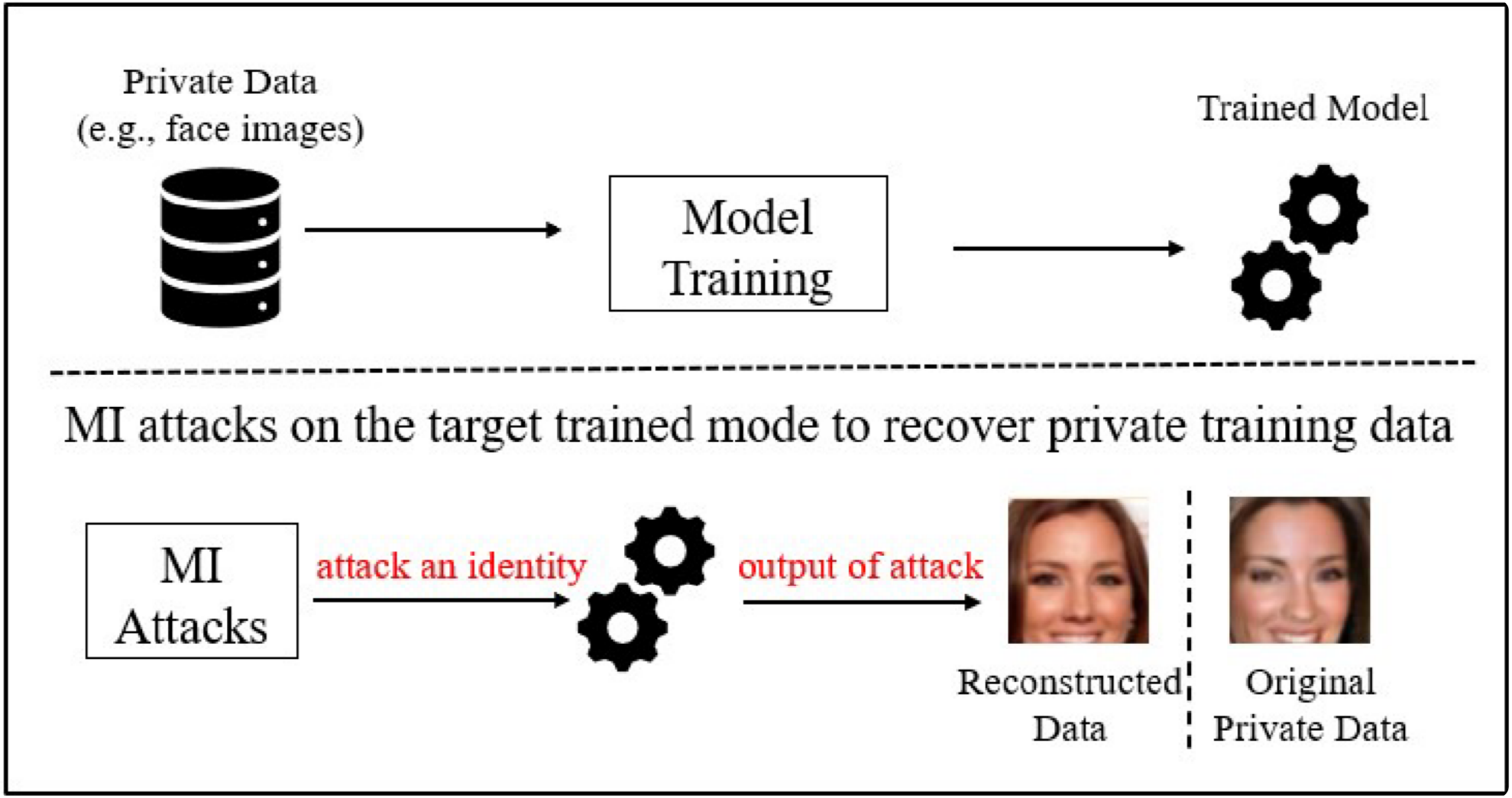

Figure: Overview of the RLB-MI attack framework.

1. Introduction

Model Inversion Attacks (MIA) exploit the correlation between a model's output and its training data to reconstruct private inputs. This project focuses on reproducing the RLB-MI attack, which formulates the model inversion problem as a reinforcement learning task in a black-box setting.

> **Original Paper:** "Reinforcement Learning-Based Black-Box Model Inversion Attacks" (Han et al., CVPR 2023)

1.1 Threat Model

- **Attacker Goal:** Reconstruct a recognizable face image of a target identity y from the classifier.
- **Access Level:** Black-box access to the target classifier (query-based, only observes output probability vector).
- **Auxiliary Knowledge:** Pre-trained Generator G trained on the same dataset (CelebA or FaceScrub) but with non-overlapping identities.

1.2 Our Contributions

- **Faithful Reproduction:** Implementation of RLB-MI with SAC agent on CelebA and FaceScrub datasets.
- **ArcFace Integration:** Use of ArcFace loss for training robust face classifiers with angular margin.
- **Temperature Scaling Analysis:** Investigation of inference-time scale factor (s=16 vs s=64) effect on attack performance.

1.3 Attack Examples

Below are examples of successful model inversion attacks. The top row shows real private training images (target identities), and the bottom row shows images reconstructed by our RL agent using only classifier queries.

Target (Private Training Data)

Recovered (Generated by Attack)

Figure 1: Model inversion attack results. Blue border = real private data, Red border = reconstructed images.

2. System Architecture

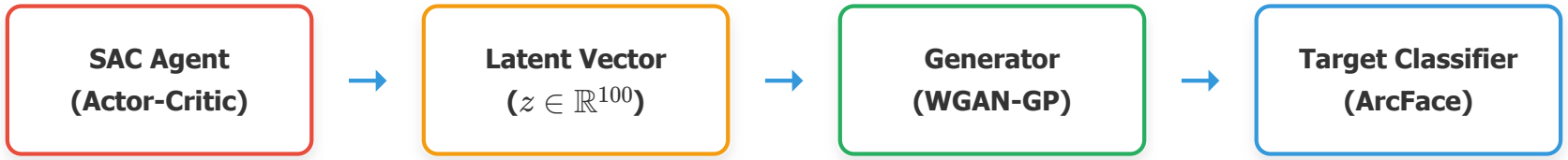

Figure 1: Attack Pipeline. The SAC agent explores the GAN latent space to find z vectors that produce images maximizing target class confidence.

2.1 ArcFace Classifier

We train face classifiers using **ArcFace (Additive Angular Margin Loss)**, which enhances discriminative power by adding an angular margin m to the target class during training:

$$L = -\log \frac{e^{s(\cos(\theta_y + m))}}{e^{s(\cos(\theta_y + m))} + \sum_{j \neq y} e^{s \cos \theta_j}}$$

- **s (scale factor):** Controls the magnitude of logits (default: 64.0 during training)
- **m (angular margin):** Adds penalty to target class angle (default: 0.5 radians ≈ 28.6°)

> **Key Insight:** During inference, ArcFace with $s = 64$ produces extremely sharp probability distributions (confidence > 99%), making the reward signal sparse. We investigate using $s = 16$ during attack to achieve more calibrated probabilities.

3. Methodology

3.1 MDP Formalization

- **State (s_t):** Current latent vector $z_t \in \mathbb{R}^{100}$
- **Action (a_t):** Next latent vector (direct jump in latent space)
- **Transition:** $z_{t+1} = \alpha z_t + (1 - \alpha) a_t$, where α is a momentum factor
- **Early Stopping:** When target classifier confidence $\geq 80\%$ for target class

3.2 Reward Function

The reward function combines three components to guide the agent:

$$R = w_1 r_1 + w_2 r_2 + w_3 r_3 \quad (w_1 = 2, w_2 = 2, w_3 = 8)$$

Component	Formula	Purpose
r_1 (State Score)	$\log P(y \mid G(z_{t+1}))$	Maximize target confidence of resulting state
r_2 (Action Score)	$\log P(y \mid G(a_t))$	Encourage selecting high-confidence actions
r_3 (Distinction Score)	$\log(P(y) - \max P(\text{others}))$	Ensure target class dominates other classes

4. Evaluation Metrics

Following the RLB-MI paper, we evaluate attacks using three complementary metrics:

1. Attack Accuracy (Top-1 / Top-5)

Definition: Percentage of generated images correctly classified as the target identity by an *independent evaluation classifier* (different from the target classifier).

- **Top-1:** Target class is the highest prediction
- **Top-5:** Target class is among the top 5 predictions

Interpretation: Higher is better. Measures attack success and transferability across classifiers.

2. KNN Distance (Feature Space)

Definition: Average L2 distance from each generated image to its K-nearest neighbors in the private training set, measured in the feature space of the evaluation classifier.

$$\text{KNN}_{\text{dist}} = \frac{1}{N} \sum \frac{1}{K} \sum \| f(\text{gen}_i) - f(\text{private}_j) \|_2$$

Interpretation: Lower is better. Indicates how similar generated images are to actual private training data.

3. FID (Fréchet Inception Distance)

Definition: Measures the distance between the distribution of generated images and private images using InceptionV3 features.

$$\text{FID} = \|\mu_{\text{gen}} - \mu_{\text{priv}}\|^2 + \text{Tr}(\Sigma_{\text{gen}} + \Sigma_{\text{priv}} - 2(\Sigma_{\text{gen}}\Sigma_{\text{priv}})^{0.5})$$

Interpretation: Lower is better. Measures perceptual quality and realism of generated faces.

5. Experimental Setup

5.1 Environment

Component	Specification
Framework	PyTorch 2.x
GPU	NVIDIA A100 (Google Colab Pro+)
Python	3.10+

5.2 Datasets

Dataset	Identities	Images	Purpose
CelebA	1,000	~30 per identity	Private training data
FaceScrub	530	~100 per identity	Private training data

Each dataset is split into **private** (for classifier training) and **public** (for GAN training) with non-overlapping identities. This simulates a realistic attack scenario where the attacker has access to similar but disjoint data.

5.3 Data Preprocessing

- **Face Detection & Alignment:** MTCNN-based face detection and 5-point landmark alignment
- **Resolution:** All images resized to 64×64 pixels
- **Normalization (Classifier):** Mean=[0.5177, 0.4284, 0.3803], Std=[0.3042, 0.2845, 0.2827]
- **Normalization (Generator):** Output range [-1, 1], converted to [0, 1] before classifier input
- **Data Augmentation:** Random horizontal flip (p=0.5) during classifier training

5.4 Model Configuration

Component	Architecture	Details
Target Classifier	VGG16 / ResNet152 + ArcFace	$s = 64, m = 0.5$, embedding_dim=512
Eval Classifier	FaceNet + ArcFace	Independent evaluation network
Generator	WGAN-GP (DCGAN-style)	$z_{\text{dim}} = 100$, output=64×64×3, trained on same dataset
RL Agent	SAC (Soft Actor-Critic)	Hidden=256×2, LR=3e-4

5.5 Attack Settings

- **Target Classes:** Top-50 most frequent identities per dataset
- **Episodes per Class:** 5,000

- **Early Stopping:** Target confidence $\geq 80\%$
- **Momentum (α):** 0.0 (direct action becomes next state)
- **Random Seed:** Fixed seed for reproducibility. Seed variation experiments were not conducted due to time constraints (~4 hours per dataset/model/50 labels).

5.6 Classifier Accuracy

The following tables show the classification accuracy of our trained target classifiers on their respective test sets:

VGG16 + ArcFace

Dataset	Top-1	Top-3	Top-5
CelebA	76.76%	84.92%	87.28%
FaceScrub	94.74%	97.37%	98.17%

ResNet-152 + ArcFace

Dataset	Top-1	Top-3	Top-5
CelebA	62.62%	72.67%	75.63%
FaceScrub	92.00%	95.49%	96.34%

Face.evoLVe (Evaluation Classifier)

Dataset	Top-1	Top-3	Top-5
CelebA	65.98%	74.63%	77.40%
FaceScrub	90.86%	94.63%	95.94%

Note: The evaluation classifier is independent from the target classifier and is used to measure attack transferability.

6. Results

6.1 Inference-Time Temperature Scaling

We investigate the effect of ArcFace scale factor during attack inference:

Scale (s)	Probability Distribution	Reward Signal	Expected Effect
s=64 (original)	Very sharp (>99%)	Sparse, binary-like	Fast convergence but local optima
s=16 (ours)	Calibrated (~70-90%)	Informative, smooth	Better exploration, higher quality

6.2 Main Results

We evaluate all combinations of datasets (CelebA, FaceScrub), target models (VGG16, ResNet152), and scale factors (s=64, s=16). **Inference-time logit scaling (s=16) consistently improves attack performance** across most configurations. Notably, using s=16 improved Top-1 accuracy by up to +18%p (FaceScrub + ResNet152) and reduced FID scores by up to 20 points, indicating better quality reconstructions.

Dataset	Target Model	Scale	Top-1 Acc (%)	Top-5 Acc (%)	KNN Dist	FID
CelebA	VGG16	s=64	16.00	32.00	1.2377	99.70
CelebA	VGG16	s=16	**22.00**	**52.00**	**1.1629**	**79.82**
CelebA	ResNet152	s=64	**18.00**	26.00	1.2164	**76.99**
CelebA	ResNet152	s=16	16.00	**30.00**	**1.1965**	78.83
FaceScrub	VGG16	s=64	22.00	34.00	1.1837	91.57
FaceScrub	VGG16	s=16	**34.00**	**50.00**	**1.1017**	**76.37**
FaceScrub	ResNet152	s=64	8.00	34.00	1.2291	76.88
FaceScrub	ResNet152	s=16	**26.00**	**44.00**	**1.1488**	**67.93**

7. Discussion

7.1 Temperature Scaling as Reward Shaping

Reducing the ArcFace scale factor from s=64 to s=16 during inference acts as **implicit reward shaping**. With s=64, the softmax output is extremely peaked, meaning small changes in the latent space produce negligible reward differences. With s=16, the probability distribution is smoother, providing more informative gradient signals to the RL agent.

7.2 Anomaly Cases: When the Attack "Succeeds" but Fails

We observed interesting cases where generated images successfully fool the target classifier with high confidence, yet **the images themselves are severely distorted or corrupted**—not even resembling a proper human face. Below are examples of such anomalies:

Figure: Anomaly cases where corrupted/blurred images still achieve high target classifier confidence.

Analysis: This phenomenon can be attributed to two factors:

- **GAN Latent Manifold Limitation:** Due to mode collapse, the GAN's latent space lacks vectors that map to proper face images for certain target identities. The RL agent navigates to regions where the generator produces corrupted outputs.
- **Classifier Over-confidence in OOD Regions:** When the agent explores latent regions that generate out-of-distribution (OOD) images, the classifier exhibits spurious high confidence due to lack of calibration in these unfamiliar regions. The agent essentially finds "adversarial" latent vectors that exploit the classifier's decision boundaries without producing meaningful images.

7.3 Limitations

- **Generator Capacity:** Attack quality depends on the GAN's ability to generate diverse, realistic faces.
- **Query Efficiency:** Thousands of queries are needed per target identity.
- **Resolution:** Current implementation uses 64×64 images; higher resolution would require more compute.

7.4 Future Work

- **Reward Shaping Exploration:** Comparing various reward shaping techniques beyond temperature scaling (e.g., potential-based shaping, curriculum learning) could yield further performance improvements.
- **Alternative RL Algorithms:** Due to the label-wise attack nature requiring separate training per target identity, we could only evaluate SAC. Comparing with other continuous control algorithms such as DDPG and TD3 would be valuable future work, though computationally expensive.
- **Modern Generative Models:** Since GANs are prone to mode collapse, exploring latent diffusion models (e.g., Stable Diffusion with VAE latent space) could improve reconstruction diversity and quality.

8. Member's Role

Member	Contributions
Daewon Kim	• Project topic selection and problem formulation • GAN (WGAN-GP) training and implementation • Environment setup and data preprocessing • Classifier training with ArcFace loss integration
Seobin Choi	• SAC agent training and hyperparameter tuning • Logit temperature scaling proposal and experiments • Report writing and documentation

9. Conclusion

We successfully reproduced the RLB-MI attack framework, demonstrating that reinforcement learning can effectively exploit black-box classifiers to reconstruct private training data. Our investigation of inference-time temperature scaling reveals that **calibrating the classifier's confidence scores** can significantly impact attack performance by providing more informative reward signals to the RL agent.

References

1. Han et al., "Reinforcement Learning-Based Black-Box Model Inversion Attacks", CVPR 2023
2. Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", CVPR 2019
3. Haarnoja et al., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep RL", ICML 2018
