Appendix

A CONFIDENCE MATCHING LOSS

We explore three key loss functions: cross-entropy loss[4, 44], max-margin loss[43] and Poincaré loss[37]. Inspired by [29], we represent the model's unnormalized scores by taking the product of the weights of the last layer and the penultimate layer activations . The cross-entropy loss aims to minimize the negative log-likelihood of the identity under the model parameters. The formulation is as follows:

$$\mathcal{L}_{ce} = -\log\left(\frac{\exp\left(p^\top w_c\right)}{\exp\left(p^\top w_c\right) + \sum_{j=1, j\neq c}\exp\left(p^\top w_j\right)}\right) \tag{6}$$

Here, p denotes the activation of the penultimate layer for sample x, and w_c represents the weight of the last layer for the c-th class in the target model M.

The maximum margin loss not only encourages maximizing the confidence score for the target class c but also emphasizes the separability of this class from others. We reformulate the maximum margin loss as follows:

$$\mathcal{L}_{mm} = -\left(p^\top w_c\right) + \max_{j\neq c}\left(p^\top w_j\right) \tag{7}$$

Here, p represents the activation of the penultimate layer for sample x. The weight of the last layer for the c-th class in the target model M is denoted as w_c. The term $p^\top w_c$ represents the unnormalized logit value for the c-th class.

Poincaré loss is a hyperbolic space embedding loss function, which can be rewritten as:

$$\mathcal{L}_p = \operatorname{arcosh}\left(1 + 2\frac{\left\|\frac{p^T w_c}{||p^T w_c||_1} - y_c\right\|_2^2}{\left(1 - \left\|\frac{p^T w_c}{||p^T w_c||_1}\right\|_2^2\right)\left(1 - ||y_c||_2^2\right)}\right) \tag{8}$$

Poincaré loss measures the distance between two vectors u and v in the hyperbolic space. $||\cdot||_2$ represents the Euclidean norm, satisfying $||u||_2 < 1$ and $||v||_2 < 1$. Here, $u = \frac{p^T w_c}{||p^T w_c||_1}$, $||\cdot||_1$ denotes the absolute value norm, v is the one-hot encoded vector for class c, denoted as y_c, and we replace 1 with 0.9999. Poincaré loss belongs to the hyperbolic distance learning paradigm, which enables measuring and comparing distances between vectors in a larger embedding space.

B EXPERIMENTAL SUPPLEMENT

B.1 Datasets

CelebA is a dataset of celebrity face attributes that contains 202,599 face pictures of 10,177 celebrity identities. For the training data of the target models (Resnet18, Resnet152, and Densenet169), we selected 1000 identities with the most number of samples, resulting in a total of 30,038 images. The FaceScrub dataset provides cropped face images of 530 identities. However, on the dataset's official website, instead of actual images, they provide download links for the dataset. All identities are used as target dataset. For Stanford.dogs, this dataset is built on top of ImageNet, which is intended for non-commercial research purposes only and provides 20,580 images of 120 dog breeds. For all datasets, the input images for the target model are resized to 224x224, while the input images for the evaluation model(Inception-V3) are resized to 299x299. The CelebA[25] dataset comprises 202,599 face pictures of 10,177 celebrity identities. In the case of training the target models (Resnet18, Resnet152, and Densenet169), we specifically selected 1,000 identities with the highest number of samples, resulting in a total of 30,038 images. Please note that the FaceScrub[27] dataset provides cropped face images of 530 identities, but the dataset's official website only offers download links instead of actual images. All identities from FaceScrub were utilized as a target dataset for our research. Regarding the Stanford.dogs[23] dataset, it is constructed on the foundation of ImageNet[7], which is exclusively intended for non-commercial research purposes. The ImageNet dataset provides 20,580 images encompassing 120 dog breeds. In our experiments, the input images for the target models were uniformly resized to dimensions of 224x224 pixels, ensuring consistency across all datasets. However, it is important to note that for the evaluation model (Inception-V3), the input images were resized to a different size of 299x299 pixels.

B.2 Publicly Available Image Prior

We downloaded the code for StyleGAN2[19] from the official source at stylegan2-ada. The AFHQ dataset consists of 16,130 high-resolution images with a resolution of 512x512 pixels. We obtained the pretrained model weights for AFHQ.dogs512[5] by using the provided link from the official code of StyleGAN2-ADA: AFHQ.dogs. It should be noted that the FFHQ[21] dataset comprises 70,000 high-quality face images. Compared to CelebA and FaceScrub, the image quality in FFHQ is significantly higher. Subsequently, we downloaded the pretrained model weights for FFHQ256 from ffhq256 and for Metfaces from Metfaces. During the attack process, for the StyleGAN2 model pretrained on FFHQ256, we performed central cropping to 200 and then resized the images to 224x224 before inputting them into the target model. For the generated images on AFHQ.dogs, we applied central cropping to 400 and then resized them to 224x224. In the comparative experiments, we followed the same cropping and resizing approach for other MIAs method.

B.3 Additional Experimental Results.

Visual Comparison of Attack Results on Facescrub. We have also visualized the attack results using the FaceScrub dataset, comparing the performance of different methods. From the visualized results in figure 6, our approach demonstrates the ability to generate synthesis images that better reflect the distinctive features of the target model's private training data.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Figure 6: We present a visual comparison of the attack results for different methods in the scenario where the $P(X_{prior}) = FFHQ$, $P(X_{target}) = Facescrub$ and the target model architecture is Resnet18. The first row shows ground truth images of target class. The second row represents PPA[37], the third row represents RLB-MI[12], and the fourth row represents Brep-MI[18]. The last row introduces our proposed method, CG-MI.