From Head to Tail: Efficient Black-box Model Inversion Attack via Long-tailed Learning

Ziang Li[1,2], Hongguang Zhang[1], Juan Wang[1]*, Meihui Chen[3], Hongxin Hu[4],
Wenzhe Yi[1], Xiaoyang Xu[1], Mengda Yang[1], Chenjun Ma[1]
[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
[2] Shanghai Innovation Institute [3] Ant Group
[4] Department of Computer Science and Engineering, University at Buffalo, SUNY
[1]{ziangli, hongguangz, jwang, wenzhey, xiaoyangx, mengday, mcj123}@whu.edu.cn
[3]chenmeihui.cmh@antgroup.com [4]hongxinh@buffalo.edu

Abstract

*Model Inversion Attacks (MIAs) aim to reconstruct private training data from models, leading to privacy leakage, particularly in facial recognition systems. Although many studies have enhanced the effectiveness of white-box MIAs, less attention has been paid to improving efficiency and utility under limited attacker capabilities. Existing black-box MIAs necessitate an impractical number of queries, incurring significant overhead. Therefore, we analyze the limitations of existing MIAs and introduce **S**urrogate **M**odel-based **I**nversion with **L**ong-tailed **E**nhancement (**SMILE**), a high-resolution oriented and query-efficient MIA for the black-box setting. We begin by analyzing the initialization of MIAs from a data distribution perspective and propose a long-tailed surrogate training method to obtain high-quality initial points. We then enhance the attack's effectiveness by employing the gradient-free black-box optimization algorithm selected by NGOpt. Our experiments show that **SMILE** outperforms existing state-of-the-art black-box MIAs while requiring only about 5% of the query overhead. Our code is available at* https://github.com/Liziang/SMILE.

1. Introduction

In recent years, deep neural networks have been widely applied in information-sensitive fields such as healthcare [48], finance [42] and image editing [60, 61], raising increasing concerns regarding privacy [18, 32, 51, 62, 64], especially training data privacy. A significant privacy threat is the model inversion attack (MIA), which aims to expose information about the training data of models. Typically, MIAs reconstruct private instances from face recognition models.

*Corresponding author.

Model Inversion Attack. [17] first proposed MIA, targeting linear regression models. Recent works have primarily focused on visual face recognition models, performing MIAs in the white-box setting, where the attacker has access to the target model parameters. GMI [69] first utilized the prior knowledge provided by GANs [19], formalizing MIA as an optimization problem. Based on this paradigm, subsequent studies have further improved the effect of reconstruction [7, 13, 40, 53, 57, 66]. Specifically, Mirror [7] utilizes StyleGAN [29] to achieve high-resolution reconstruction results on open-source pre-trained models. In black-box MIA, attackers can obtain the model output. RLBMI [21] combines reinforcement learning, and Mirror [7] employs genetic algorithms to solve this optimization problem. [28, 39] focus on the more challenging setting where attackers can only access hard labels.

Long-tailed Learning. The long-tail distribution is typically reflected in the fact that a few individuals contribute the most, dominating the dataset as head classes, while most classes, called tail classes, have few data samples [63, 67]. Models trained on long-tail distributions often exhibit bias towards head classes, leading to suboptimal performance on tail classes. Common mitigation solutions such as resampling [12, 16], reweighting [15, 25], margin modifications [10, 36], and ensemble learning [9, 59] are employed to improve long-tailed recognition performance.

In this paper, we propose an efficient black-box MIA, named **S**urrogate **M**odel-based **I**nversion with **L**ong-tailed **E**nhancement (*SMILE*). We analyze the limitations of existing black-box MIAs from two perspectives: the targeted datasets and the number of queries. To address these limitations, we follow the principle that *high-quality initial points are crucial for optimization problems* and conduct detailed data analysis and visualization of the logits output by the

target model. By leveraging the proposed **Long-tailed surrogate training** method, we construct high-quality local surrogate models with an extremely limited number of queries, which significantly reduces the complexity of the subsequent black-box optimization process. In summary, our contributions are as follows:

- We perform a fine-grained data analysis on the output logits from open-source pre-trained models. We find that the number of samples across classes exhibits an extreme long-tail distribution and discuss possible mitigation strategies.
- We model the training of surrogate models as a long-tailed recognition problem. By utilizing the proposed long-tailed surrogate training method, we obtain high-quality surrogate models with extremely limited queries. This provides advantageous initial points for subsequent optimization processes.
- We propose *SMILE*, an efficient black-box MIA, which integrates long-tailed surrogate training and the black-box optimization algorithm selected by NGOpt. Experiments demonstrate that *SMILE* surpasses existing state-of-the-art black-box MIAs, while requiring only about 5% of the query overhead.

2. Motivation

Although MIAs have been extensively studied, state-of-the-art black-box MIAs still face significant limitations that seriously affect their practicability and effectiveness.

2.1. The targeted datasets

One primary limitation stems from the characteristics of the targeted private training datasets. Similarly to most MIAs, including GMI [69], KEDMI [13], VMI [57], PLGMI [66], LOMMA [40], BREPMI [28], and LOKT [39], the datasets targeted by RLBMI [21] feature low-resolution images and a limited number of IDs. These datasets generally undergo alignment, clipping, and resizing to a uniform size of 64×64. Examples include CelebA [34], which contains 1,000 private IDs out of 10,177 total IDs; FaceScrub [38], which contains 200 private IDs out of 530 total IDs; and PubFig83 [45], with 50 private IDs out of 83 total IDs. Moreover, in the main experiments, GAN models trained on samples from the remaining IDs serve as image priors. This typical setting reduces the authenticity and complexity of the MIA, because it simplifies the process of MIAs via GAN latent space optimization to disclose private facial features. Therefore, we follow Mirror's setting [7] where:

- The target models are pre-trained models available online.
- The number of IDs in the private training data is massive.
- The synthetic data from pre-trained GAN models and the private training data originate from different distributions.

These alleviate the above limitation, as detailed in Tab. 2.

2.2. The number of queries

Another limitation arises from the number of queries to the black-box target model. In black-box attacks, query cost is a crucial metric, second only to the attack's success rate. One significant drawback of current black-box MIAs is the monetary cost caused by the large number of queries. For instance, utilizing a commercial face recognition API costs \$0.001 or more per call [1, 3, 6]. If an attacker requires over $100K$ queries to compromise the facial privacy of a target ID, the expense would amount to \$100. Furthermore, during an attack on a fixed ID, the optimization process of MIAs tends to generate numerous intermediate results with similar facial features. It is easy for the cloud service provider hosting the API to detect such large-scale and similar access requests and then interrupt the black-box MIAs.

Regrettably, current black-box MIAs rely on executing a large number of queries. A general MIA process consists of latent vector sampling for initialization ①, iterative optimization for feature search ②, and an optional post-processing step ③ [46]. For a selected target model, Step ① is required only once, while Step ② and ③ repeat for each targeted ID. As illustrated in Tab. 1, for Mirror-b [7], $100K$ samples are required for initialization. Each attack subsequently incurs an additional query overhead of $20K$. For RLBMI [21], although Step ① is not needed, it involves a query overhead of $80K$ in Step ②. In Tab. 1, we also detail the query costs of white-box MIAs. Although the number of queries is less critical for white-box MIAs where attackers can obtain model parameters and run locally, a high number of initialization and optimization still increases the time and resource costs.

MIAs	Type	Step ①	Step ②	Step ③	$N=1$	$N=10$	$N=50$
Mirror-w [7]	White-box	100K	20K	—	120K	300K	1.1M
PPA [53]	White-box	5K	14K	5K	24K	195K	0.995M
RLBMI [21]	Black-box	0K	80K	—	80K	800K	4M
Mirror-b [7]	Black-box	100K	20K	—	120K	300K	1.1M
SMILE (ours)	Black-box	2.5K	1K	—	3.5K	12.5K	52.5K

Table 1. **Statistics of the number of queries required by MIAs.** N is the number of targeted IDs. The query overhead of our method is only about 5% compared to the overhead of SOTA black-box MIAs.

2.3. Our solution

From the perspective of minimizing query overhead, one might easily conclude that an efficient black-box MIA should incorporate:

1) An initialization conducive to the optimization.
2) An efficient black-box optimization algorithm.
3) A one-shot attack mechanism for a specific target ID.

Specifically, for Step ①, more private information regarding the target ID should be mined. For Step ②, a black-box optimization algorithm suitable for searching the GAN latent space is required, preferably without Step ③. To this end, we first feed GAN-synthesized facial samples into pre-trained facial recognition models and perform data

MIAs	Dateset (N_{priv}) & Model Architecture & Resolution					Image Priors
[13, 21, 28] [39, 40, 57] [66, 69]	PubFig83 (50) [45]	FaceScrub (200) [38]		CelebA (1000) [34]		IID Data FFHQ [29]
	VGG16 [52], ResNet152 [22], Facenet64 [14] 64 × 64					
PPA [53]		FaceScrub (530) [38]		CelebA (1000) [34]		FFHQ [29] MetFaces [30]
		ResNet [22], ResNeSt [68], DenseNet [26] Series 224 × 224				
Mirror [7] Ours	VGGFace (2622) [43]		VGGFace2 (9131) [11]		CASIA (10575) [65]	CelebA [34] FFHQ [29]
	VGG16 [5] 224 × 224	VGG16BN [5] 224 × 224	ResNet50 [5] 224 × 224	InceptionV1 [2] 160 × 160	InceptionV1 [2] 160 × 160	SphereFace [4] 112 × 196

Table 2. **The datasets and models involved in MIAs.** The number of privacy IDs highlighted in red. Most MIAs focus on low-resolution scenes and utilizing independently and identically distributed (IID) data as image priors, with only partial experiments involving shifted priors from FFHQ. We target a large number of private IDs, using image priors from different distributions, which implies more difficult MIA.

analysis on the logits. From this, we introduce a surrogate training method based on long-tailed learning. By attacking the local surrogate models, Step ① of our method provides high-quality initialization points for Step ② under a strict query budget. Then, in Step ②, we apply the black-box optimization algorithm selector NGOpt [33, 37], provided by Nevergrad [47]. NGOpt automatically selects a suitable gradient-free black-box optimization as the solver based on high-level problem information. Based on the above, our solution achieves superior black-box MIA effects under an extremely constrained query budget, as detailed in Tab. 1.

3. Methodology

3.1. Threat model

MIAs aim to uncover private training data based on the target model M_t, which is typically modeled as an optimization problem with auxiliary priors. The auxiliary priors, derived from the public dataset $\mathcal{D}_{pub}$, share the same or similar data distribution with the private dataset $\mathcal{D}_{priv}$. To narrow the search space and facilitate the optimization process, the prior knowledge provided by $\mathcal{D}_{pub}$ is compressed into a generative model G, typically an attacker-self-trained or pre-trained GAN model.

In the workflow of a GAN-based MIA, the attacker initially samples a set of latent vectors $\mathcal{Z}_{init} = \{z_1, \ldots, z_{N_{Step-1}}\}$ from the standard Gaussian distribution and employs the GAN model to generate an initial sample pool $\mathcal{X}_{init} = G(\mathcal{Z}_{init}) = \{x_1, \ldots, x_{N_{Step-1}}\}$, N_{Step-1} is the number of samples in Step ①. These samples are then fed into the M_t. Based on the output logits $\mathcal{Y}_{init} = M_t(\mathcal{X}_{init}) = \{y_1, \ldots, y_{N_{Step-1}}\}$, the attacker selects the candidate $x \sim G(z)$ that optimally corresponds to the target ID $c \in \mathcal{C}_{target} = \{1, \ldots, N_{target}\}$ — typically the one closer to the target individual in perceptual distance. The candidate z serves as an initialization and proceeds to the subsequent optimization process, formalized as:

$$\min_{z \in \mathcal{Z}_{init}} \mathcal{L}_{id}(M_t(G(z)), c) + \lambda \mathcal{R}(G, z). \quad (1)$$

$\mathcal{R}$ is regularization and the optimization process iterates N_{Step-2} times to minimize the loss, and finally z^* is obtained. The attacker takes $x^* = G(z^*)$ as the MIA result on c. In the case of black-box MIA [7, 21], the attacker can query M_t but cannot access its parameters, so that Eq. (1) cannot be optimized via backpropagation, which is the primary reason for the extensive number of queries required in black-box MIAs. We assume attackers know the application domain of the target model, but are not aware of its training method or model architecture. For practicality, in our main experiments, both the black-box target models and the GAN models are from open-source, pre-trained models. Additionally, in our setting, the attacker is able to obtain a model pre-trained on an arbitrary face dataset. This is feasible, as there are numerous such pre-trained models available on the Internet. We also further evaluate models trained for MIA defense purposes.

3.2. Data analysis on logits

For training surrogate models M_s, the typical paradigm involves using input-output pairs $(\mathcal{X}_{init}, \mathcal{Y}_{init})$ derived from the M_t. The M_t functions as the teacher model, and in a knowledge distillation manner [23], the information from the $\mathcal{D}_{priv}$ is transferred to a locally pre-trained model which acts as the student model. This process is formalized as:

$$\min_{M_s} \mathbb{E}_{(x,y)\sim(\mathcal{X}_{init}, \mathcal{Y}_{init})} \left[D_{KL}(q(y|x) \| p(y|x; M_s)) \right], \quad (2)$$

where D_{KL} is Kullback-Leibler divergence [31], and the feature extractor of M_s is initialized from M_{pre}, which is a recognition model pretrained on $\mathcal{D}_{pre}$, a dataset whose distribution differs from $\mathcal{D}_{priv}$.

However, we find that even with a large number (up to $10K$) of input-output pairs, the black-box MI attacker is still unable to effectively construct a high-quality surrogate model. The specific performance is the excessively low test accuracy of the M_s on the $\mathcal{D}_{priv}$, as shown in Tab. 4. This means that the resulting surrogate model obtained in this way is not sufficiently similar in predictive behavior to the target model, as the test accuracy serves as a crucial metric for the effect of knowledge distillation [20] and model stealing [27, 41]. This further leads to the fact that performing

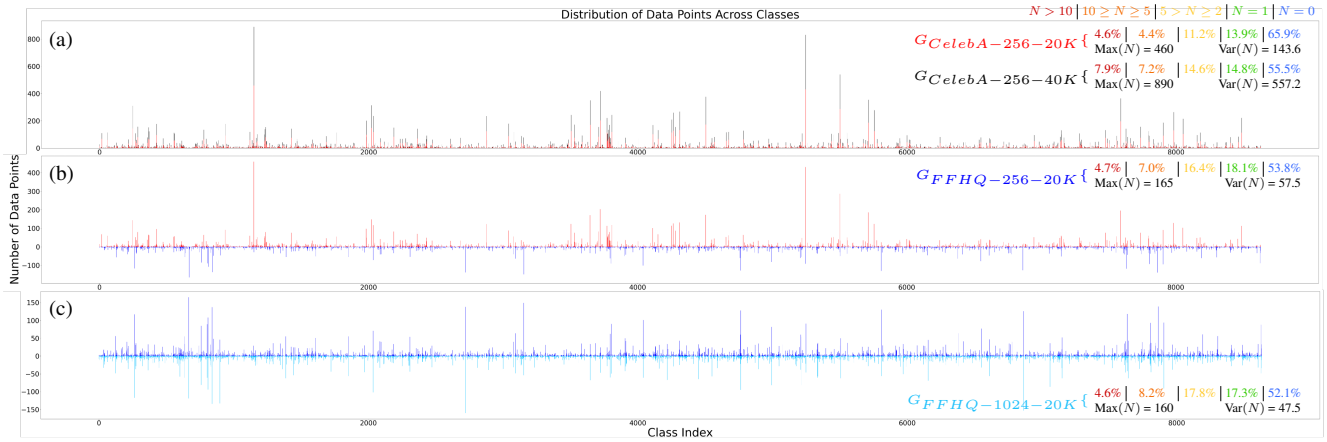


Figure 1. **Visualization of sample distribution across classes.** N denotes the number of samples in a single label. The colors of the bar chart indicate the GAN models used for sampling. From **(a)**, it can be seen that with the same image prior, doubling the sampling causes a significant increase in the variance of N, but the proportion of $N = 0$ only slightly decreases. This indicates that the long-tail distribution is not mitigated. It is also evident that classes with extremely high sample counts at $20K$ sampling remain so at $40K$ sampling, demonstrating the inefficiency of blindly increasing the sampling size. In **(b)**, it can be observed that the bar chart obtained from FFHQ sampling is flatter than that of CelebA, indicating that using a more diverse image prior can effectively mitigate the long-tail distribution, specifically reflected in the smaller variance of N. In **(c)**, using a stronger GAN under the same image priors provides only minimal improvement for the long-tail distribution, as evidenced by a slight reduction in variance. Interestingly, classes that occupy more samples under the weaker GAN sampling often also appear in the strong GAN sampling, manifesting as symmetry in the bar charts above and below. However, charts derived from different prior samplings tend to be asymmetrical, as shown in **(b)**. This suggests that the mitigation of long-tail distributions by image priors primarily depends on the data distribution, rather than the advancement of the generative model.

white-box MIAs on the local substitution model can only obtain extremely poor attack results. Next, we analyze the causes of this low accuracy.

We use a VGGFace2 pre-trained ResNet50 model [5] as the M_t and G trained on CelebA as the auxiliary image prior for our example. By analyzing $20K$ logits, we find that the number of samples allocated to each class shows a very extreme long-tail distribution. Among the 8631 IDs, only 9% of the classes have no less than 5 samples, while 65.9% of the classes have no samples assigned at all. Notably, some classes occupy far more samples than the average, as shown in Fig. 1(a). We believe that the reason for this phenomenon is that CelebA, as an auxiliary dataset, has a significant difference from the target dataset VGGFace2 in the fine-grained distribution of facial feature combinations. Despite the extensive sample size, the intersection of the two covers only a portion of the private IDs. We further conjecture that blindly increasing the sample size has a limited contribution to improving the quality of surrogate models. This is because, for generative models, a large number of samples from different batches will have very close distributions despite the randomness. In simple terms, doubling the sample size will not substantially improve the extreme long-tail distribution. While there may be a slight increase in the number of intersecting IDs, the majority of the samples still fall into the already heavily populated classes, as shown in Fig. 1(a). More sampled data is redundant for the training of surrogate models, but greatly increases the query burden of black-box MIAs.

One possible mitigation strategy involves using an auxiliary $\mathcal{D}_{pub}$ that encompasses a richer combination of facial features, thereby ensuring greater diversity, to achieve more overlap with the private IDs. Ideally, when the auxiliary $\mathcal{D}_{pub}$ is diverse enough and its data distribution overlaps with the $\mathcal{D}_{priv}$ sufficiently, uniformly sampling from a high-quality generative model could effectively cover private IDs. Specifically, when utilizing the FFHQ pre-trained GAN model (FFHQ includes vastly more variation than CelebA [29]), with an equivalent sample size, FFHQ achieves a broader intersection and exhibits reduced variance in the number of data points per class. However, the issue of extreme long-tail distribution persists, as illustrated in Fig. 1(b). Even when using a more powerful GAN model, the mitigation of the long-tail problem remains limited, as shown in Fig. 1(c). Considering that attackers are often unable to select the most suitable GAN model, our solution follows: Without relying on increasing sample sizes or enhancing auxiliary priors, within a limited sampling budget, we aim to extract as much information as possible from the private training data in the context of extreme long-tail distribution, to obtain higher-quality surrogate models.

3.3. SMILE

Based on the analysis presented, we propose **S**urrogate **M**odel-based **I**nversion with **L**ong-tailed **E**nhancement (*SMILE*), a two-step efficient black-box MIA. In the first step, we introduce a long-tailed surrogate training method specifically tailored for the MIA by integrating various long-tailed learning techniques. In the second step, using the results of attacking the local surrogate model as initialization, we apply a gradient-free optimization algorithm to enhance the effectiveness of the attack.

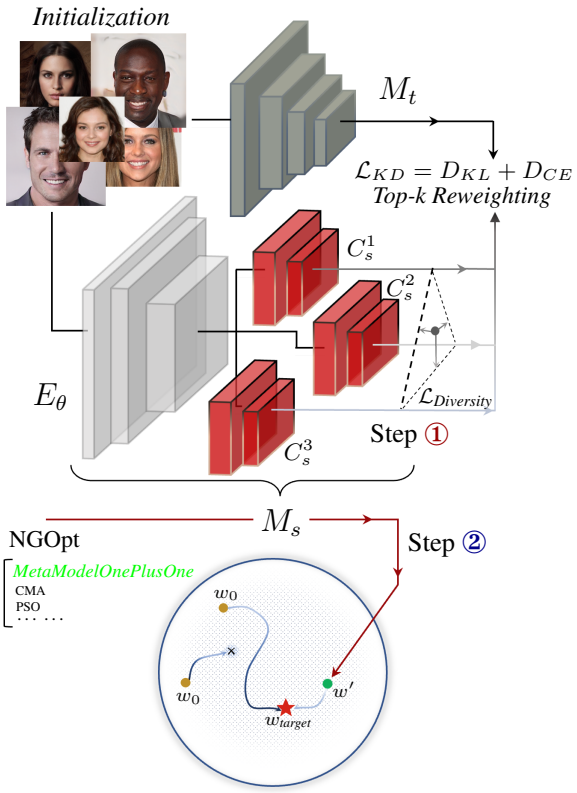


Figure 2. The overall architecture of ***SMILE***.

Long-tailed surrogate training Considering the attacker's access to only a very limited dataset of $(\mathcal{X}_{init}, \mathcal{Y}_{init})$ pairs, we update merely the parameters of the final two layers of M_s during training. Because fine-tuning M_{pre} with a too small dataset, which is misaligned with the private dataset, would lead to a catastrophic collapse in its feature extraction capabilities. Appendix A.1 corroborates this view. To address the challenges posed by the extremely long-tail distribution of the training data, our strategy aims to maximize the extraction of information from the tail class while ensuring comprehensive learning of the head class information. Inspired by RIDE [59], we design an ensemble method with Distribution-aware diversity loss [59].

The pre-trained M_{pre} consists of a feature extractor $f_\theta = \{f_{\theta 1}, \ldots, f_{\theta n}\}$ with n layers and a classifier C_{pre}. For training M_s, we freeze the parameters of the first $n-1$ layers, using them as a shared backbone $E_\theta = \{f_{\theta 1}, \ldots, f_{\theta n-1}\}$. We then construct N new ensemble classifiers $C_s = \{f_{\theta n}, C_t'\}$, where C_t' is randomly initialized but maintains the same output dimension as M_t. The inference of M_s can be formalized as follows:

$$M_s(x) = \frac{1}{N}\sum_{i=1}^{N} C_s^i(E_\theta(x)). \tag{3}$$

All these N classifiers C_s are trained together using a distribution-aware diversity loss $\mathcal{L}_{Diversity}$ and a distillation loss $\mathcal{L}_{KD}$. The calculation of $\mathcal{L}_{KD}$ is independent within the ensemble models, ensuring that each model complements the others [59], as formalized below:

$$\begin{aligned}\mathcal{L}_{KD} = \sum_{i=1}^{N}[D_{KL}(q(y|x)||p(y|x; C_s^i \circ E_\theta)), \\ + \ \alpha_{ce} * D_{CE}(y_{pseudo}, p(y|x; C_s^i \circ E_\theta))],\end{aligned} \tag{4}$$

where $(x, y) \in (\mathcal{X}_{init}, \mathcal{Y}_{init})$, α_{ce} is the hyperparameter that balances D_{KL} with D_{CE}. The ensemble comprises N models, set to 3 (ablation of N is in Appendix A.2). The pseudo-hard label y_{pseudo} is employed, where labels are built according to $\mathcal{Y}_{init}$ by selecting the class with the highest probability. The guidance of pseudo-hard labels ensures efficient learning of head classes by the ensemble models.

$\mathcal{L}_{Diversity}$ serves as a regularization term to promote complementary decisions from multiple models of the ensemble. The implementation strategy involves maximizing the Kullback-Leibler divergence between the prediction probabilities of each individual model and the ensemble's average prediction probabilities. To ensure training stability, we minimize the sum of the reciprocals as:

$$\mathcal{L}_{Diversity} = \sum_{i=1}^{N} \frac{1}{D_{KL}(p_i(y|x; C_s^i \circ E_\theta)||p_{avg}(y|x; M_s))}, \tag{5}$$

where $p_i(y|x; C_s^i \circ E_\theta)$ denotes the predicted probability distribution of the i-th model. To further mitigate the issue of long-tail distribution, we adopt a different approach from the common reweighting method [10, 15]. Considering that the Top-1 labels of the samples only cover a subset of all classes, and these classes are sparsely distributed across the entire set, we propose a *Top-k Reweighting* strategy. We adjust the weights based on the frequency of each class appearing in the Top-k prediction probabilities, formalized as follows: For $y_n \in \{y_1, \ldots, y_{Step-1}\}$, find the class index of the corresponding Top-k largest logits $T_n = \{t_{n1}, \ldots, t_{nk}\}$, where t_{nk} denotes the class index corresponding to the k-th largest logits of y_n. For each class $c \in \mathcal{C}_{target} = \{1, \ldots, N_{target}\}$, count the total number of its occurrences in the Top-k logits of all samples:

$$\text{count}(c) = \sum_{n=1}^{N_{Step-1}} \sum_{j=1}^{k} \delta(t_{nj} = c), \delta(A) = \begin{cases} 1, & if A \\ 0, & else. \end{cases} \tag{6}$$

Set $\beta = 0.9$, and for each class c, calculate the weights:

$$\text{weights}(c) = \frac{1-\beta}{1-\beta^{\text{count}(c)+1}}, \tag{7}$$

$$W = \text{weights}(t_{n1}). \tag{8}$$

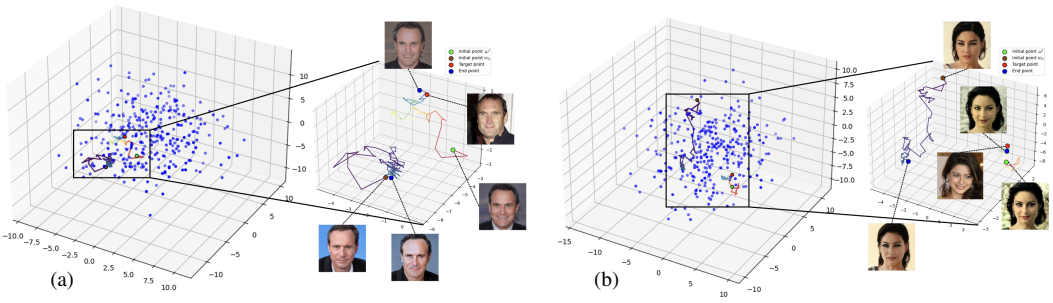


Figure 3. **We visualized the optimization processes under different initializations.** w_0 is the sample with the highest confidence for the target ID in the sample pool, and w' is the sample obtained from the white-box MIA on M_s. As shown in **(a)**, long-tailed surrogate training provides a more helpful initial point for black-box optimization, allowing it to approach the target point in very few iterative steps while avoiding local optima. **(b)** shows that long-tailed surrogate training fully captures the information of this ID, thereby providing a high-quality initial point close to the target point.

In summary, the total loss for our proposed Long-tailed surrogate training can be described as follows:

$$\mathcal{L}_{total} = W * \mathcal{L}_{KD} + \alpha_{diversity} * \mathcal{L}_{Diversity}. \quad (9)$$

It is worth noting that although our total loss function has several hyperparameters, we empirically chose a fixed setting throughout our experiments ($\alpha_{ce} = 0.15, \alpha_{diversity} = 10$, *Top-10 Reweight*), which illustrates the robustness of our method. We perform Mirror-w [7] on M_s and send the results $w' \in \mathcal{W}'_{init}$ to the second step, formalized as:

$$\min_{w \in \mathcal{W}_{init}} \mathcal{L}_{id}(M_s(G(w)), c), \quad (10)$$

$$w_i = \text{LeakyReLU}(Clip(\text{LeakyReLU}^{-1}(w_i), \mu \pm \sigma)). \quad (11)$$

where $\mathcal{W}_{init} = G_{mapping}(\mathcal{Z}_{init})$. Eq. (11) serves as a regularization term, performing $\mathcal{P}$ space pruning on w after each update. The upper and lower bounds are derived from the mean and variance of $\mathcal{W}_{init}$ in $\mathcal{P}$ space.

Gradient-free black-box optimization Considering the limitation that the initial sampling covers less than 50% of the private IDs. Although the results from the first step can be classified by M_s to the target ID, the corresponding facial features are often not accurately aligned with the target privacy instance. To improve the confidence of attack results in M_t, we aim for more accurate feature matching and generalized attack outcomes.

We employ the black-box optimization algorithm selector NGOpt. Given the latent feature dimension of 512 and a black-box query budget of 2500 (the reason for selecting this budget is presented in Appendix A.3), NGOpt automatically determines the appropriate black-box optimization algorithm *MetaModelOnePlusOne* [8, 47]. We conduct black-box attacks on M_t with it, leveraging the initialization from the first step. The black-box optimization process is particularly challenging compared to the white-box setting due to the lack of gradient information, which makes it harder to search for private features solely based on the output. To address this challenge, we further release the latent vector search capability of GANs. In the optimization process, we adopt the $\mathcal{P}$ space clipping method [7, 70] and enlarge the clipping interval, formalized as follows:

$$\min_{w' \in \mathcal{W}'_{init}} \mathcal{L}_{id}(M_t(G(w')), c), \quad (12)$$

$$w'_i = \text{LeakyReLU}(Clip(\text{LeakyReLU}^{-1}(w'_i), \mu \pm k * \sigma)). \quad (13)$$

After the optimization, the attacker obtains the final $x^* = G(w^*)$. In the experiments, we set $k = 1.7$ empirically. The overall architecture of *SMILE* is shown in Fig. 2. We further visualize the attack process of *SMILE* by Principal Component Analysis (PCA) in Fig. 3.

4. Experiments

In the experimental section, we evaluate the current SOTA MIAs targeting high-resolution scenarios, including Mirror [7], PPA [53], and the SOTA black-box MIA RLBMI [21]. We take the results of SOTA white-box MIAs as an upper bound for our attack's effectiveness, given that having access to the parameters significantly reduces the complexity of the optimization problem. We do not consider other MIAs targeting low-resolution scenarios, as their distortion in high-resolution scenarios is predictable.

4.1. Experimental setup

For fairness, we adhered to the default configurations of other MIAs, including hyperparameters, GAN types, and the number of queries, as specified in their papers or official open-source repositories. Details are in Appendix B.

Dataset. We focused on facial recognition tasks, specifically the pre-trained face classification models for VGGFace [43], VGGFace2 [11], and CASIA [65] datasets. For image priors, $\mathcal{D}_{pub}$ and $\mathcal{D}_{priv}$ in all experiments originated from distinct distributions to maintain the practicality of MIA. Details are included in Tab. 2 and Appendix B.

Models. Following Mirror [7], we utilize GANs [50] pre-trained on CelebA [34] and FFHQ [29]. For the setting of the pre-trained model that initializes the surrogate model,

$\mathcal{D}_{priv}$	VGGFace						CASIA					
M_t	VGG16		VGG16BN		InceptionV1				SphereFace			
Method	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓
Mirror-w	81.63	89.80	77.55	93.88	57.14	75.51	446.96	403.66	61.22	77.55	379.51	388.24
PPA	97.96	100.0	100.0	100.0	77.55	83.67	385.72	347.79	57.14	77.55	363.42	375.14
Mirror-b	59.18	73.47	57.14	79.59	28.57	44.90	601.762	559.36	16.33	30.61	425.63	440.86
RLBMI	71.43	83.67	71.43	91.84	16.33	36.73	597.84	556.45	0.0	6.12	575.32	578.50
ResNet50*	70.07±1.92	87.76±1.67	71.43±0.0	78.23±1.92	37.66±0.85	53.41±1.26	459.27	419.53	40.41±2.65	58.89±3.36	442.19	452.48
InceptionV1*	71.43±1.67	89.12±0.96	63.27±1.67	72.79±0.96	45.84±4.44	63.00±0.38	435.79	403.83	39.46±1.93	58.50±0.96	444.19	445.67
InceptionV3	72.11±1.92	87.76±1.67	64.63±0.96	82.31±0.96	43.54±2.54	73.47±4.41	525.96	477.83	51.70±1.92	61.91±1.93	392.27	404.40
MobileNetV2	71.43±1.67	82.99±1.93	68.71±0.96	85.03±1.93	47.62±0.96	71.43±2.89	624.43	560.87	48.30±2.54	65.99±4.19	398.85	407.20
EfficientNetB0	70.75±1.92	84.35±2.55	71.21±1.95	79.45±1.5	44.90±3.33	65.99±0.96	500.48	458.35	46.26±2.54	63.26±2.89	417.64	424.17
Swin-T	71.90±1.14	83.53±1.84	72.11±0.96	80.95±1.92	45.58±0.96	65.99±3.47	496.31	453.34	48.30±2.54	64.63±2.55	396.13	405.47

$\mathcal{D}_{priv}$	VGGFace2							
M_t	ResNet50				InceptionV1			
Method	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓
Mirror-w	63.27	79.59	313.33	280.97	51.02	61.22	2878.88	2938.51
PPA	100.0	100.0	161.54	150.97	93.88	100.0	1933.23	2028.02
Mirror-b	24.49	38.78	369.55	332.88	26.53	46.94	2714.71	2736.29
RLBMI	42.86	65.31	375.58	328.12	40.82	65.31	2934.68	2963.05
InceptionV1*	44.22±2.54	61.90±5.36	339.16	306.98	41.50±0.96	63.95±0.96	2790.31	2760.00
EfficientNetB0	39.46±0.96	49.66±1.92	392.75	351.45	50.34±0.96	63.27±1.67	2667.81	2672.07

$\mathcal{D}_{priv}$	VGGFace2							
M_t	ResNet50				InceptionV1			
Method	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓
Mirror-w	81.63	97.96	227.06	208.90	81.63	85.71	2044.31	2078.71
PPA	97.96	97.96	149.35	144.49	100.0	100.0	1663.03	1776.31
Mirror-b	44.90	67.35	289.38	264.83	55.10	81.63	1895.33	1970.20
RLBMI	63.27	89.80	323.07	281.20	71.43	85.71	2551.42	2589.16
InceptionV1*	68.71±1.92	93.20±0.96	262.56	237.75	89.80±1.67	94.56±0.96	1779.18	1899.43
EfficientNetB0	74.15±2.54	89.12±0.96	257.65	239.22	79.59±1.67	89.80±0.0	1805.03	1882.68

$\mathcal{D}_{priv}$	VGGFace						CASIA					
M_t	VGG16		VGG16BN		InceptionV1				SphereFace			
Method	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓	Acc@1↑	Acc@5↑	KNN Dist↓	Feat Dist↓
Mirror-w	93.88	100.0	87.76	100.0	77.55	87.76	465.96	415.81	73.47	91.84	319.26	330.82
PPA	100.0	100.0	97.96	100.0	85.71	95.92	456.28	404.90	59.18	83.67	352.20	361.18
Mirror-b	59.18	79.59	61.22	75.51	46.94	71.43	693.57	635.81	22.45	40.82	393.98	400.51
RLBMI	67.35	91.84	61.22	77.55	53.06	71.43	598.79	576.03	16.33	24.49	520.43	519.73
ResNet50*	80.27±4.19	91.16±1.92	56.46±2.54	78.91±1.92	49.36±5.46	68.47±3.57	575.16	523.97	38.39±4.53	55.49±3.20	406.14	413.89
InceptionV1*	80.95±3.47	91.16±0.96	63.95±0.96	80.95±0.96	58.50±1.92	80.27±0.96	459.21	421.65	48.30±3.47	65.99±2.54	400.65	410.08
InceptionV3	72.79±1.92	89.12±0.96	62.59±0.97	87.08±0.97	64.63±2.55	80.95±2.54	649.73	581.97	53.74±0.96	67.35±1.66	370.07	376.73
MobileNetV2	79.59±4.41	90.48±0.96	75.36±3.15	91.78±1.60	62.59±0.97	76.87±2.54	673.40	607.23	58.50±5.85	75.51±1.67	346.74	362.32
EfficientNetB0	72.79±2.54	89.80±1.66	70.75±3.85	80.95±3.47	59.86±2.55	73.47±3.33	590.52	534.32	55.78±1.92	70.07±0.96	356.78	359.19
Swin-T	80.95±1.92	88.44±1.93	68.71±0.96	85.71±1.67	59.58±1.76	87.01±4.12	578.46	515.81	47.94±0.83	72.60±0.86	362.93	374.42

Table 3. **Performance report of MIAs on different private datasets and target model architectures.** The red / blue indicates the optimal **black-box** MIA results when the image prior is CelebA / FFHQ. The bolded **Arch** indicates the surrogate model architecture used by ***SMILE***. * refers to the surrogate model initialized with a pre-trained face recognition model on the Internet. **Swin-T** is the abbreviation for **Swin Transformer**.

we ensure that their training data come from distributions different from the private training data. Specifically, for VGGFace and CASIA, we use pre-trained ResNet50 [5] and InceptionV1 [2] from VGGFace2; for VGGFace2, InceptionV1 [2] pre-trained on CASIA is used. To verify the robustness of the proposed long-tailed surrogate training method for M_{pre}, we self-train models across multiple architectures [35, 49, 55, 56], details are in Appendix B.

Evaluation metrics. Following previous research [28, 40, 69], we select three metrics: Attack Accuracy (**Attack Acc**) : Following Mirror [7], we employ two models pre-trained on the same dataset, each serving as the evaluation model for the other, and report the **Acc@1** and **Acc@5**; K-Nearest Neighbors Distance (**KNN Dist**); Feature Distance (**Feat Dist**), details are in Appendix B.

4.2. Experimental results

Tab. 3 presents a comprehensive overview of *SMILE*'s attack results, illustrating that our method is SOTA for black-box MIA under most conditions, spanning various private datasets and diverse targeted pretrained models. The employment of multiple surrogate model architectures and image priors underscores *SMILE*'s robustness, qualitative results are in Fig. 6 and Appendix G. It can be observed that the effectiveness of *SMILE* varies with different surrogate model architectures. This variation is due to different surrogate models' capability to extract features and their alignment with private datasets, affecting their adaptation to the target domain and consequently impacting the long-tailed surrogate training. To verify the robustness of our method, we fixed a uniform set of hyperparameters in the training of surrogate models, resulting in certain quality differences among different architectures. We consider this acceptable.

For RLBMI, which performs suboptimally, we observe inflated metrics. This inflation occurs because under a large number of queries, some results are optimized into adversarial samples with transferability. Consequently, certain outcomes from RLBMI appear significantly distorted, yet they receive high confidence scores from the evaluation model, as shown in Fig. 4. This phenomenon is more obvious when targeting the VGG dataset, where a relatively small number of classes are easier to generate adversarial examples. In contrast, $\mathcal{P}$ space clipping effectively mitigates this problem.

Figure 4. **Some results of RLBMI on VGG.** The image is seriously distorted, but the evaluation model has high confidence in them.

Ablation study. We incrementally add loss terms and observe their impact on the performance of the surrogate model. The *Base* is the model obtained by applying D_{KL} and only updating the final classifier layer. Tab. 4 shows that

$\mathcal{D}_{priv}$	VGGFace2, InceptionV1							
$\mathcal{D}_{pub}$	CelebA				FFHQ			
Method	*Base* D_{KL}	***SMILE*** D_{KL}+$D_{Diversity}$	***SMILE*** ··· +D_{CE}	***SMILE*** ··· +*Top-k Reweighting*	*Base* D_{KL}	***SMILE*** D_{KL}+$D_{Diversity}$	***SMILE*** ··· +D_{CE}	***SMILE*** ··· +*Top-k Reweighting*
InceptionV1*	5.00/11.74	12.94/25.11	16.21/29.59	21.84/39.29	12.07/26.20	25.42/43.99	28.67/47.60	36.54/58.90
EfficientNetB0	4.32/11.51	6.47/14.90	6.85/15.40	9.80/20.74	7.77/18.61	11.57/24.67	11.66/24.06	17.46/33.22
InceptionV1*	6.36/14.77	10.37/21.9	16.31/30.94	20.63/37.44	15.39/31.01	24.13/44.07	25.96/44.97	40.17/61.00
EfficientNetB0	6.19/15.18	9.46/20.67	10.15/21.01	15.45/29.81	9.91/22.63	16.40/33.05	16.15/31.60	24.06/42.86
InceptionV1*	8.53/18.21	19.28/35.47	22.78/39.34	24.87/42.08	19.14/36.43	36.51/57.61	40.38/60.68	45.12/64.68
EfficientNetB0	7.91/18.60	13.56/27.98	16.49/31.14	20.91/37.94	12.80/27.33	25.39/44.90	23.86/42.78	31.54/51.59

$\mathcal{D}_{priv}$	VGGFace2, ResNet50							
$\mathcal{D}_{pub}$	CelebA				FFHQ			
Method	*Base* D_{KL}	***SMILE*** D_{KL}+$D_{Diversity}$	***SMILE*** ··· +D_{CE}	***SMILE*** ··· +*Top-k Reweighting*	*Base* D_{KL}	***SMILE*** D_{KL}+$D_{Diversity}$	***SMILE*** ··· +D_{CE}	***SMILE*** ··· +*Top-k Reweighting*
InceptionV1*	2.58/6.88	5.25/11.92	6.18/14.40	8.91/19.40	6.78/16.74	12.85/26.91	14.00/27.86	20.45/38.98
EfficientNetB0	2.03/5.82	2.49/6.69	2.51/6.56	3.08/7.76	4.05/11.44	4.87/12.81	5.01/12.69	7.01/16.55
InceptionV1*	3.48/8.64	4.84/11.15	5.76/13.06	9.16/19.67	9.34/21.16	14.35/30.17	17.84/34.12	23.78/42.76
EfficientNetB0	2.97/8.15	3.67/9.03	3.71/9.28	4.75/11.59	5.67/15.45	7.40/17.47	7.78/18.06	11.57/25.07
InceptionV1*	5.18/11.70	8.88/19.06	11.45/23.31	14.27/28.28	12.73/26.19	24.64/44.80	24.79/43.80	32.29/52.51
EfficientNetB0	4.14/10.83	5.19/12.51	6.25/14.62	7.13/16.81	8.45/19.87	12.50/27.03	13.87/28.61	17.51/34.26

Table 4. **Ablation study of loss terms.** We set the sample sizes to $2.5K$, $5K$, and $10K$. For each setting, the loss terms are added sequentially from left to right, with green indicating performance improvement and red indicating performance decline. It can be observed that the performance of the surrogate models generally shows an upward trend and is significantly better than *Base*. This suggests that each loss term contributes.

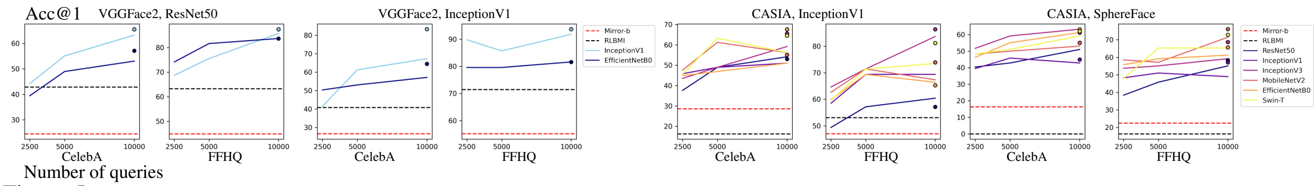


Figure 5. **Setting initial sample sizes to $2.5K$, $5K$, and $10K$ reveals a gradual increase in the effectiveness of *SMILE*.** This trend suggests that while a larger sample size does not directly address the long-tail distribution, it provides more privacy information. The surrogate model, relying on more information, can offer higher quality initial points, making the subsequent optimization process easier. Additionally, by extending the optimization steps from $1K$ to $2K$, an enhancement in efficacy is observable (the **dots** in the graph). This demonstrates *SMILE*'s potential when expanding the query budget.

in various settings, the increase in the loss terms of *SMILE* almost invariably leads to positive effects, and the quality of the obtained surrogate models significantly surpasses that of *Base*. We further increase the initial sampling size. Fig. 5 indicates that this enhancement improves *SMILE*'s performance, highlighting its potential for large-scale query.

Defense. We further evaluate various defense mechanisms, including BiDO [44], MID [58], LS [54], and TL [24]. As shown in Tab. 5, although there is a decrease in performance, our method still significantly outperforms existing black-box MIAs. We also find that defenses have a much greater effect on black-box MIAs than on white-box MIAs. We will focus on developing black-box MIAs that are insensitive to defenses as our future research interest. Details are in Appendix C.

Other experiments. We further use art face [30] as image priors, which have a significant distribution difference from private data, as shown in Appendix A.4. We discuss the challenges faced by label-only MIA under pre-trained models and large-scale private ID settings, and introduce the concept of Attack-sensitive ID, which includes General Attack-sensitive ID and Dataset-specific Attack-sensitive ID. Details are provided in Appendix A.5.

			PPA		Mirror-b		RLBMI		SMILE	
Defenses	Hyperparameters	Acc	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑
BiDO	0.006, 0.06	92.20 (2.27↓)	93.88	95.92	12.24	16.33	8.16	16.32	22.45	40.82
	0.03, 0.3	91.57 (1.63↓)	55.11	75.51	6.12	14.29	8.16	22.44	22.49	36.73
MID	0.005	91.31 (3.16↓)	93.88	100.00	8.16	24.49	6.12	22.44	14.29	26.53
	0.005	89.50 (3.70↓)	83.67	93.87	4.08	4.08	18.36	24.48	6.12	10.20
LS	-0.001	92.40 (2.07↓)	87.76	95.92	8.16	10.20	4.08	14.28	30.61	38.78
	-0.0005	92.48 (0.72↓)	65.31	77.55	10.20	14.29	8.16	12.24	24.49	34.69
TL	Block 4	93.82 (0.65↓)	81.63	95.92	6.12	10.20	14.28	16.32	20.41	36.73
	Block 3	91.98 (1.22↓)	34.69	59.18	2.04	6.12	4.08	6.12	10.20	24.49

Table 5. **Performance of MIAs under defenses.** The $\mathcal{D}_{priv}$ is VGGFace2. Red refers to MobileNetV2, and blue refers to Swin Transformer.

5. Conclusion

In this paper, we introduce *SMILE*, an efficient black-box MIA. By combining long-tailed surrogate training and gradient-free black-box optimization, *SMILE* outperforms existing black-box MIAs while requiring about 5% of the query overhead. Experiments on various datasets and model architectures demonstrate its robustness.

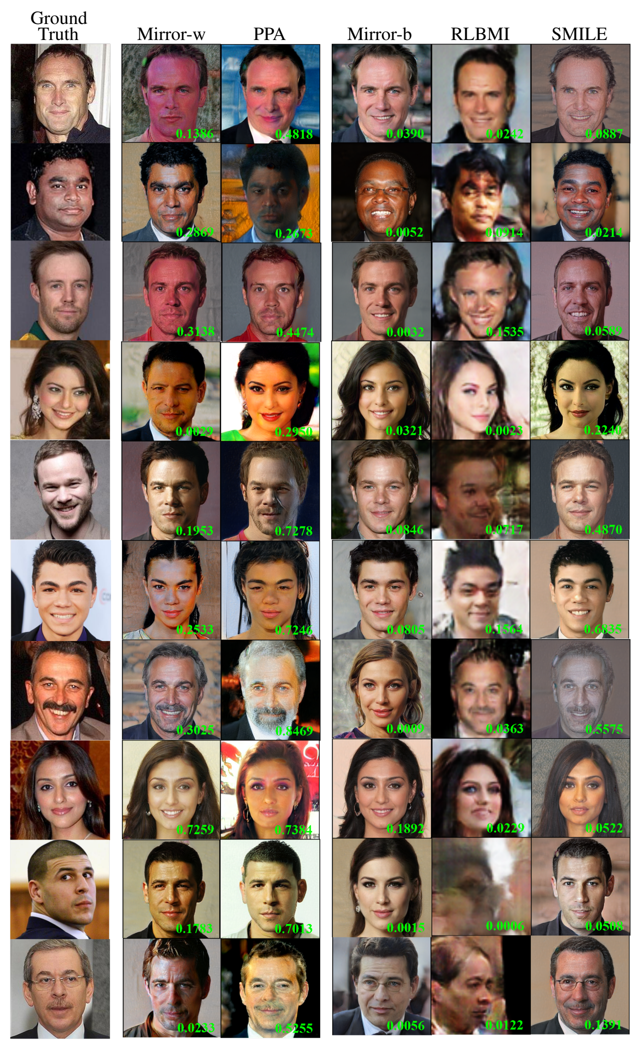


Figure 6. **Results of MIAs.** M_t is ResNet50 and CelebA for image prior. The M_s of *SMILE* is InceptionV1.

References

[1] Amazon rekognition. https://aws.amazon.com/rekognition/pricing/. 2

[2] Inception resnet (v1) models in pytorch. https://github.com/timesler/facenet-pytorch. 3, 7

[3] Microsoft face api. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/face-api/. 2

[4] Sphereface model in pytorch. https://github.com/clcarwin/sphereface. 3

[5] Vggface/vggface2 models in pytorch. https://www.robots.ox.ac.uk/~albanie/pytorch-models.html. 3, 4, 7

[6] Face++ services. https://www.faceplusplus.com/v2/pricing/. 2

[7] Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. Mirror: Model inversion for deep learning network with high fidelity. In *Proceedings of the 29th Network and Distributed System Security Symposium*, 2022. 1, 2, 3, 6, 7

[8] Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies–a comprehensive introduction. *Natural computing*, 1:3–52, 2002. 6

[9] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 112–121, 2021. 1

[10] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1, 5

[11] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 3, 6

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 1

[13] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16178–16187, 2021. 1, 2, 3

[14] Yu Cheng, Jian Zhao, Zhecan Wang, Yan Xu, Karlekar Jayashree, Shengmei Shen, and Jiashi Feng. Know you at one glance: A compact vector representation for low-shot learning. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1924–1932, 2017. 3

[15] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 1, 5

[16] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, 2003. 1

[17] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium*, pages 17–32, 2014. 1

[18] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020. 1

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[20] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 3

[21] Gyojin Han, Jaehyun Choi, Haeil Lee, and Junmo Kim. Reinforcement learning-based black-box model inversion attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 20504–20513, 2023. 1, 2, 3, 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[23] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[24] Sy-Tuyen Ho, Koh Jun Hao, Keshigeyan Chandrasegaran, Ngoc-Bao Nguyen, and Ngai-Man Cheung. Model inversion robustness: Can transfer learning help? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12183–12193, 2024. 8

[25] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 1

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3

[27] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium*, pages 1345–1362, 2020. 3

[28] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 15045–15053, 2022. 1, 2, 3, 7

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 3, 4, 6

[30] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 3, 8

[31] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951. 3

[32] Ziang Li, Mengda Yang, Yaxin Liu, Juan Wang, Hongxin Hu, Wenzhe Yi, and Xiaoyang Xu. Gan you see me? enhanced data reconstruction attacks against split inference. *Advances in neural information processing systems*, 36: 54554–54566, 2023. 1

[33] Jialin Liu, Antoine Moreau, Mike Preuss, Jeremy Rapin, Baptiste Roziere, Fabien Teytaud, and Olivier Teytaud. Versatile black-box optimization. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 620–628, 2020. 3

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 3, 6

[35] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 7

[36] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 1

[37] Laurent Meunier, Herilalaina Rakotoarison, Pak Kan Wong, Baptiste Roziere, Jeremy Rapin, Olivier Teytaud, Antoine Moreau, and Carola Doerr. Black-box optimization revisited: Improving algorithm selection wizards through massive benchmarking. *IEEE Transactions on Evolutionary Computation*, 26(3):490–500, 2021. 3

[38] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing*, pages 343–347. IEEE, 2014. 2, 3

[39] Bao-Ngoc Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Man Cheung. Label-only model inversion attacks via knowledge transfer. *Advances in neural information processing systems*, 36, 2024. 1, 2, 3

[40] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking model inversion attacks against deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 16384–16393, 2023. 1, 2, 3, 7

[41] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*, 55 (14s):1–41, 2023. 3

[42] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *Applied soft computing*, 93:106384, 2020. 1

[43] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. 3, 6

[44] Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1358–1367, 2022. 8

[45] Nicolas Pinto, Zak Stone, Todd Zickler, and David Cox. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In *CVPR 2011 workshops*, pages 35–42. IEEE, 2011. 2, 3

[46] Yixiang Qiu, Hongyao Yu, Hao Fang, Wenbo Yu, Bin Chen, Xuan Wang, Shu-Tao Xia, and Ke Xu. Mibench: A comprehensive benchmark for model inversion attack and defense. *arXiv preprint arXiv:2410.05159*, 2024. 2

[47] Jérémy Rapin and Olivier Teytaud. Nevergrad-a gradient-free optimization platform, 2018. 3, 6

[48] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020. 1

[49] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[50] Yujun Shen, Yinghao Xu, Ceyuan Yang, Jiapeng Zhu, and Bolei Zhou. Genforce. https://github.com/genforce/genforce, 2020. 6

[51] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy*, pages 3–18. IEEE, 2017. 1

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[53] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting. Plug & play attacks: Towards robust and flexible model inversion attacks. *arXiv preprint arXiv:2201.12179*, 2022. 1, 2, 3, 6

[54] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Be careful what you smooth for: Label smoothing can be a privacy shield but also a catalyst for model inversion attacks. *arXiv preprint arXiv:2310.06549*, 2023. 8

[55] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[56] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7

[57] Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. *Advances in neural information processing systems*, 34:9706–9719, 2021. 1, 2, 3

[58] Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11666–11673, 2021. 8

[59] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020. 1, 5

[60] Yibin Wang, Weizhong Zhang, and Cheng Jin. Magicface: Training-free universal-style human image customized synthesis. *arXiv preprint arXiv:2408.07433*, 2024. 1

[61] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. High-fidelity person-centric subject-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7675–7684, 2024. 1

[62] Xiaoyang Xu, Mengda Yang, Wenzhe Yi, Ziang Li, Juan Wang, Hongxin Hu, Yong Zhuang, and Yaxin Liu. A stealthy wrongdoer: Feature-oriented reconstruction attack against split learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12130–12139, 2024. 1

[63] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022. 1

[64] Mengda Yang, Ziang Li, Juan Wang, Hongxin Hu, Ao Ren, Xiaoyang Xu, and Wenzhe Yi. Measuring data reconstruction defenses in collaborative inference systems. *Advances in neural information processing systems*, 35:12855–12867, 2022. 1

[65] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3, 6

[66] Xiaojian Yuan, Kejiang Chen, Jie Zhang, Weiming Zhang, Nenghai Yu, and Yang Zhang. Pseudo label-guided model inversion attack via conditional generative adversarial network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3349–3357, 2023. 1, 2, 3

[67] Xinli Yue, Ningping Mou, Qian Wang, and Lingchen Zhao. Revisiting adversarial training under long-tailed distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 24492–24501, 2024. 1

[68] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2746, 2022. 3

[69] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 253–261, 2020. 1, 2, 3, 7

[70] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 6

From Head to Tail: Efficient Black-box Model Inversion Attack via Long-tailed Learning

Supplementary Material

A. Additional experimental results

A.1. Full parameter fine-tuning weakens the extraction ability

M_t	VGGFce2, ResNet50			
Image Priors	CelebA		FFHQ	
Method	*Base*	*Base*$_{full}$	*Base*	*Base*$_{full}$
InceptionV1*	2.58/6.88	0.42/1.47	6.78/16.74	1.09/3.48
InceptionV1*	3.48/8.64	0.41/1.47	9.34/21.16	1.37/3.95
InceptionV1*	5.18/11.70	0.82/2.63	12.73/26.19	3.40/8.45

M_t	VGGFce2, InceptionV1			
Image Priors	CelebA		FFHQ	
Method	*Base*	*Base*$_{full}$	*Base*	*Base*$_{full}$
InceptionV1*	5.00/11.74	0.98/2.96	12.07/26.20	3.11/7.78
InceptionV1*	6.36/14.77	1.47/4.12	15.39/31.01	3.36/8.48
InceptionV1*	8.53/18.21	3.40/8.21	19.14/36.43	7.96/16.92

Table 1. **Acc@1 / Acc@5 for fine-tuning only the classifier (*Base*) and fine-tuning all parameters (*Base*$_{full}$).** The number of samples is set to $2.5K$, $5K$, and $10K$. * refers to the surrogate model initialized with a pre-trained face recognition model obtained from the Internet. It can be seen that surrogate models obtained through full parameter fine-tuning suffer a severe drop in accuracy on the private dataset. This indicates that when the sample size is small, fine-tuning all parameters severely degrades the feature extraction capability of the surrogate model.

A.2. Ablation study on the number of models in the ensemble

M_t	VGGFace2, ResNet50				
Image Priors	CelebA				
Method	*SMILE* $N=3$	*SMILE* $N=4$	*SMILE* $N=5$	*SMILE* $N=10$	*SMILE* $N=50$
InceptionV1*	8.91/19.40	9.95/21.29	9.54/20.96	9.59/20.81	8.22/19.61
EfficientNetB0	3.08/7.76	3.10/8.09	3.16/8.09	3.06/7.83	3.21/8.26
InceptionV1*	9.16/19.67	8.65/19.43	9.95/21.59	9.79/20.89	9.33/19.88
EfficientNetB0	4.75/11.59	5.04/12.18	5.08/12.23	5.23/12.57	5.32/12.62
InceptionV1*	14.27/28.28	14.09/27.90	14.36/28.46	14.93/28.89	12.84/27.09
EfficientNetB0	7.13/16.81	7.43/17.35	7.73/17.79	7.69/17.77	8.75/19.37

M_t	VGGFace2, ResNet50				
Image Priors	FFHQ				
Method	*SMILE* $N=3$	*SMILE* $N=4$	*SMILE* $N=5$	*SMILE* $N=10$	*SMILE* $N=50$
InceptionV1*	20.45/38.98	20.79/39.36	21.08/39.21	21.02/39.30	20.79/40.41
EfficientNetB0	7.01/16.55	7.00/16.73	6.98/16.57	7.04/16.90	7.18/16.97
InceptionV1*	23.78/42.76	25.65/45.57	23.81/42.57	24.79/43.82	25.04/44.10
EfficientNetB0	11.57/25.07	11.94/25.70	11.41/24.93	12.02/25.76	11.99/25.86
InceptionV1*	32.29/52.51	32.47/52.66	34.07/54.39	33.19/53.70	33.43/53.77
EfficientNetB0	17.51/34.26	17.96/35.01	18.13/35.23	18.95/36.45	19.80/37.82

M_t	VGGFace2, InceptionV1				
Image Priors	CelebA				
Method	*SMILE* $N=3$	*SMILE* $N=4$	*SMILE* $N=5$	*SMILE* $N=10$	*SMILE* $N=50$
InceptionV1*	21.84/39.29	20.41/37.96	20.20/36.05	22.11/39.03	20.05/37.80
EfficientNetB0	9.80/20.74	8.97/19.11	10.00/21.11	9.81/21.07	9.74/20.90
InceptionV1*	20.63/37.44	20.88/37.15	21.32/37.90	21.88/39.66	21.14/38.82
EfficientNetB0	15.45/29.81	15.36/29.44	15.53/30.09	15.34/29.24	15.71/30.25
InceptionV1*	24.87/42.08	24.35/41.55	24.92/42.57	25.13/42.48	24.70/41.98
EfficientNetB0	20.91/37.94	20.68/37.58	21.10/37.89	22.70/39.58	24.06/42.06

M_t	VGGFce2, InceptionV1				
Image Priors	FFHQ				
Method	*SMILE* $N=3$	*SMILE* $N=4$	*SMILE* $N=5$	*SMILE* $N=10$	*SMILE* $N=50$
InceptionV1*	36.54/58.90	38.06/58.09	38.50/58.61	37.05/59.09	34.27/55.75
EfficientNetB0	17.46/33.22	17.10/33.05	17.07/32.59	17.97/34.02	18.30/34.36
InceptionV1*	40.17/61.00	37.91/58.81	39.31/59.47	40.94/61.52	39.86/60.96
EfficientNetB0	24.06/42.86	24.43/43.13	24.56/43.54	25.59/44.71	26.81/45.61
InceptionV1*	45.12/64.68	44.78/64.42	44.59/64.20	46.32/65.82	44.44/64.75
EfficientNetB0	31.54/51.59	32.61/52.50	32.43/52.03	33.19/52/90	34.39/54.77

Table 2. **Acc@1 / Acc@5 for surrogate models with N models in the ensemble.** The number of samples is set to $2.5K$, $5K$, and $10K$. * refers to the surrogate model initialized with a pre-trained face recognition model obtained from the Internet. We highlighted the highest-quality surrogate models under a specific setting in red and the lowest-quality surrogate models in blue. As observed, when the sample size is $2.5K$, setting $N = 5$ is more likely to yield higher-quality surrogate models. Additionally, as the sample size increases, the quality of the surrogate models shows a positive correlation with the value of N. While using $N = 5$ is more likely to produce better surrogate models with 2500 samples, selecting $N = 3$ in our main experiments is reasonable. This is because, in the context of black-box MIAs, attackers should not have prior knowledge of the optimal value of N. The main experiments demonstrate that even with a suboptimal N, *SMILE* can still achieve desirable attack performance. Furthermore, we recommend increasing N as the sample size grows to better account for the greater amount of private information.

A.3. Why 2500 queries

Please refer to Appendix A.5 and Tab. 6.

A.4. Art face as the image prior

Please refer Appendix G.

A.5. Challenges in the label-only setting

Label-only is a challenging setting for MIAs because the information available to attackers is extremely limited, and the issue is intensified by the large-scale private ID settings. Existing label-only MIAs require at least one sample corresponding to the target ID to be collected before an attack can be launched, serving as an initial point for subsequent optimization processes [21] or for training a T-ACGAN [39]. This is feasible when the number of private IDs is relatively small (e.g., 50/200/530/1000), but for scenarios with a large number of private IDs, even sampling up to $40K$ samples, over 50% of private IDs still do not receive any samples (as shown in Fig. 1), meaning attackers cannot obtain any information about these IDs. Existing MIAs cannot be effectively launched, and *SMILE* faces the same issue, as an initial sampling of $2.5K$ covers only a

Image priors	CelebA&FFHQ		Examples
Sampling size	Intersection size	Proportion	The indexes
$40K$	63	21.0%	[5248, 3803, 7906, 2035, 3646, 3722, 5810, 7149, 365, 5503, 273, 3795, 2086, 8488, 3772, 7800, 4551, 7148, 3791, 553]
$20K$	61	20.33%	[5248, 3803, 2035, 7906, 3646, 3722, 5810, 7149, 365, 5503, 273, 3795, 8488, 2086, 7800, 3078, 2472, 3772, 7148, 2309]
$10K$	60	20.0%	[5248, 3803, 7906, 3646, 3722, 2035, 5810, 7149, 5503, 273, 365, 3795, 8488, 2309, 3772, 7800, 2086, 4551, 553, 2472]
$5K$	58	19.33%	[5248, 3803, 3646, 3722, 7906, 2035, 5810, 5503, 3795, 273, 7149, 3772, 8488, 4551, 2309, 2086, 365, 7800, 4506, 3791]
$2.5K$	47	15.66%	[5248, 3803, 3646, 3722, 2035, 7906, 5810, 5503, 273, 7149, 3772, 3795, 2086, 8488, 365, 1234, 7800, 8407, 3078, 553]
$1K$	46	15.33%	[5248, 2035, 3722, 3646, 7906, 273, 3784, 1234, 5758, 5503, 7149, 8407, 8488, 8193, 8227, 3772, 4113, 1427, 4896, 5710]
$0.5K$	32	10.66%	[5248, 2035, 3646, 5810, 3722, 8488, 7149, 8193, 8407, 1234, 1427, 7641, 365, 7042, 934, 4679, 7886, 741, 4979, 884]

Table 3. We set M_t as ResNet50 pre-trained on VGGFace2. We calculate the intersection of the top 300 dataset-specific attack-sensitive IDs obtained using CelebA and FFHQ as image priors across various sampling sizes, which serve as the general attack-sensitive IDs. It can be observed that a part of IDs are simultaneously easy to be covered under different image priors, and we consider them to be the most attack-sensitive instances in the privacy dataset. Taking the top 20 general attack-sensitive IDs across various sampling sizes as an example, it can be found that the top 20 general vulnerable IDs generally emerge when the sampling size is $2.5K$.

small portion of private IDs, as shown in Tab. 4. Therefore, in our setup, launching a label-only MIA for each ID is unfeasible, and we do not wish to increase the number of queries to millions like existing label-only MIAs [21, 39]. In the label-only setting, we propose a new objective: To compromise as many private IDs as possible with as few queries as necessary. We introduce the concept of **Attack-sensitive ID**, which includes **General Attack-sensitive ID** and **Dataset-specific Attack-sensitive ID**. Attack-sensitive IDs, in the context of label-only MIAs, refer to IDs that receive more samples at initialization, meaning that these IDs are relatively more exposed to attackers. For a specific ID, having access to more samples provides attackers with the opportunity to either directly expose private information or obtain better initial points that are beneficial for subsequent optimization. Dataset-specific attack-sensitive ID refers to attack-sensitive IDs under specific image priors, which are easier to attack under this prior, details in Tab. 6. General attack-sensitive ID includes IDs that are vulnerable under various image priors and represents the intersection of different dataset-specific attack-sensitive IDs, as Tab. 3.

We do not perform iterative optimization and only use long-tail surrogate training. We perform white-box MIAs on local surrogate models targeting dataset-specific attack-sensitive IDs, with results presented in Tab. 7.

B. Details of experimental setup

Datasets & Models. Following [7], we target face recognition models that are pre-trained on VGGFace, VGGFace2, and CASIA datasets, all of which are obtained from the Internet. This means that our attack does not need for the private training datasets themselves. However, to demonstrate the quality of the surrogate models (measured as Acc@1 and Acc@5 on the private dataset), extend the pre-trained model architectures used for initializing the surrogate models, and compute KNN Dist and Feat Dist, we need access to the private training datasets. We obtain them from this[1][2], and it is noted that these data are not high-quality original data and have not been strictly aligned with the pre-trained models, leading to lower accuracy of the pre-trained models on the test dataset. We randomly sample 10% from each dataset as the test dataset to evaluate the pre-trained and surrogate models. Using the remaining 90% as training data, we pretrain multiple models on various architectures as initializations for surrogate models. It is important to note that in the experiments, the pre-trained models used to initialize the surrogate models and the target pre-trained models are pre-trained on data from different distributions. The details of the models are shown in Tab. 5. The pre-trained GAN models used in the experiments to generate high-resolution images are from this[3].

Image priors	CelebA				
Sampling size	$N > 10$	$10 \geq N \geq 5$	$5 > N \geq 2$	$N = 1$	$N = 0$
$2.5K$	0.36%	0.94%	3.26%	7.31%	88.13%
$5K$	1.04%	1.74%	5.07%	9.73%	82.42%
$10K$	2.28%	3.07%	7.61%	12.16%	74.88%
$20K$	4.55%	4.41%	11.19%	13.99%	65.86%

Image priors	FFHQ				
Sampling size	$N > 10$	$10 \geq N \geq 5$	$5 > N \geq 2$	$N = 1$	$N = 0$
$2.5K$	0.22%	0.65%	4.36%	10.28%	84.49%
$5K$	0.61%	1.91%	7.61%	14.13%	75.74%
$10K$	1.90%	4.08%	12.17%	15.88%	65.97%
$20K$	4.67%	7.06%	16.39%	18.05%	53.83%

Table 4. We set M_t as ResNet50 pre-trained on VGGFace2. N denotes the number of samples in single label. As observed, limited sampling leads to the majority of IDs not receiving samples. This means that attacking each ID is not feasible.

Hyperparameters of MIAs. The settings for the number of queries are shown in Tab. 1. For Mirror-w, the optimizer is adam with a learning rate of 0.2; For PPA, the optimizer is adam with a learning rate of 0.005; For

[1] https://www.kaggle.com/datasets/hearfool/vggface2

[2] https://www.kaggle.com/datasets/debarghamitraroy/casia-webface

[3] https://github.com/genforce/genforce

Role	Architecture	Training dataset	Input resolution	Classes	Source	Report Acc@1	Test Acc@1	Epoch	Batch size	Optimizer	Learning rate
M_t	VGG16	VGGFace	224*224	2622	[5]	97.22	-	-	-	-	-
M_t	VGG16BN	VGGFace	224*224	2622	[5]	96.29	-	-	-	-	-
M_t/M_s	ResNet50	VGGFace2	224*224	8631	[5]	99.88	96.99	-	-	-	-
M_t/M_s	InceptionV1	VGGFace2	160*160	8631	[2]	99.65	93.70	-	-	-	-
M_s	InceptionV3	VGGFace2	342*342	8631	-	-	95.04	20	64	adam	0.001
M_s	MobileNetV2	VGGFace2	224*224	8631	-	-	94.47	20	128	adam	0.001
M_s	EfficientNetB0	VGGFace2	256*256	8631	-	-	96.69	20	128	adam	0.001
M_s	Swin-T	VGGFace2	260*260	8631	-	-	93.21	6	20	adam	0.001
M_t/M_s	InceptionV1	CASIA	160*160	10575	[2]	99.05	87.31	-	-	-	-
M_t	SphereFace	CASIA	112*96	10575	[4]	99.22	-	-	-	-	-
M_s	EfficientNetB0	CASIA	256*256	10575	-	-	91.24	60	128	adam	0.001

Table 5. **Details of the models.** We are unable to obtain the VGGFace dataset, so we do not test the accuracy, as well as the KNN Dist or Feat Dist in our experiments.

Image priors	CelebA		FFHQ	
Sampling size	Intersection size	Proportion	Intersection size	Proportion
$40K$	2000	100%	2000	100%
$20K$	1671	83.55%	1717	85.85%
$10K$	1526	76.3%	1500	75.0%
$5K$	1272	63.6%	1382	69.1%
$2.5K$	1005	50.25%	1110	55.5%
$1K$	736	36.8%	814	40.7%
$0.5K$	630	31.5%	664	33.2%

Table 6. We set M_t as ResNet50 pre-trained on VGGFace2. For various sampling sizes, we calculated the top 2000 IDs containing the highest number of samples. We use the top 2000 IDs from a sample size of $40K$ as an approximation of the top 2000 dataset-specific attack-sensitive IDs with the utilized image prior, since it involves substantial sampling. We then analyze the intersection size and proportion of these top 2,000 IDs at reduced sampling sizes with the top 2000 IDs at $40K$ sampling. Notably, when the sampling size is reduced to $2.5K$, the intersection still exceeds 50%. This indicates that even with a significant reduction in sampling size, $2.5K$ samples can still provide a decent approximation of the image prior. We believe this is sufficient for black-box MIAs, which is why we set the sampling size to $2.5K$. Similarly, when the sampling size is reduced to $0.5K$, the top 2000 IDs still overlap by more than 30% with those under the $40K$ sampling size. Thus, dataset-specific attack-sensitive IDs begin to manifest even at lower sampling sizes, giving attackers the potential to identify these vulnerable IDs earlier, causing the acceleration of privacy leakage.

RLBMI, we directly use the settings in their open source code. For *SMILE*, in all experiments, the hyperparameters for long-tailed surrogate training are uniformly set to $\alpha_{ce} = 0.15, \alpha_{diversity} = 10$, *Top-10 Reweight*, the hyperparameters for gradient-free black-box optimization are set to $k = 1.7$, the optimizer is adam with a learning rate of 0.2. All temperatures T used for distillation with KL divergence are set to 0.5.

Evaluation metrics. Following Mirror [7], we employ two models pre-trained on the same dataset, each serving as the evaluation model for the other, and report the **Acc@1** and **Acc@5**; K-Nearest Neighbors Distance (KNN Dist) measures the shortest distance in the feature space between the reconstructed image and the private images of the target ID; Feature Distance (Feat Dist) measures the distance between the feature of the reconstructed image and the average feature of the target ID's private images. The feature distance is the l_2 distance between the outputs from the penultimate layer of the evaluation model. We attack the first 49 IDs of all datasets, and our main experiment in Tab. 3 is repeated 3 times.

$\mathcal{D}_{priv}$	VGGFace2			
Image Priors	CelebA			
M_t	ResNet50		InceptionV1	
Method	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑
InceptionV1*	22.45	44.90	16.33	38.78
EfficientNetB0	16.33	30.61	18.37	32.65
InceptionV1*	22.45	40.82	26.53	61.22
EfficientNetB0	24.49	48.98	26.53	61.22
InceptionV1*	16.33	30.61	28.57	59.18
EfficientNetB0	20.41	34.69	36.73	57.14
InceptionV1*	24.49	40.82	32.65	65.31
EfficientNetB0	24.49	55.10	38.78	69.39

$\mathcal{D}_{priv}$	VGGFace2			
Image Priors	FFHQ			
M_t	ResNet50		InceptionV1	
Method	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑
InceptionV1*	24.49	53.06	38.78	59.18
EfficientNetB0	26.53	36.73	38.78	63.27
InceptionV1*	28.57	55.10	34.69	69.18
EfficientNetB0	28.57	42.86	34.69	61.22
InceptionV1*	24.49	51.02	46.94	65.31
EfficientNetB0	34.69	55.10	59.18	77.55
InceptionV1*	30.61	57.14	55.10	77.55
EfficientNetB0	36.73	63.27	44.90	73.47

Table 7. The number of samples is set to $2.5K$, $5K$, $10K$, and $20K$. * refers to the surrogate model initialized with a pre-trained face recognition model obtained from the Internet. Targeting dataset-specific attack-sensitive IDs, long-tailed surrogate training can effectively obtain private information even in the label-only setting and very limited number of queries.

C. Details of defenses

We implement the defenses on MobileNetV2 and Swin Transformer pre-trained on VGGFace2, and the hyperparameters and experimental results are shown in Tab. 5 and Tab. 8. We note that the gradient-free black-box optimization process is severely disturbed when attacking the model under MID defense. We believe that this is caused by the random noise introduced by MID during inference, which makes it difficult for the black-box optimization process to converge. Therefore, for MID, we only use the white-box attack results on the surrogate models.

			PPA		Mirror-b		RLBMI		SMILE	
Defenses	Hyperparameters	Acc	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑
BiDO	0.006, 0.06	92.20 (2.27↓)	89.80	97.96	14.29	46.94	24.49	36.73	30.61	57.14
	0.03, 0.3	91.57 (1.63↓)	63.27	91.84	10.20	26.53	18.37	26.53	26.53	57.14
MID	0.005	91.31 (3.16↓)	100.00	100.00	16.33	44.90	40.81	57.14	28.57	44.90
	0.005	89.50 (3.70↓)	91.83	97.95	6.12	14.29	26.53	42.86	6.12	12.24
LS	-0.001	92.40 (2.07↓)	83.67	93.88	14.29	38.78	20.41	32.65	20.41	38.78
	-0.0005	92.48 (0.72↓)	59.18	67.35	14.29	28.57	22.45	36.73	30.61	48.98
TL	Block 4	93.82 (0.65↓)	81.63	93.88	14.29	40.82	28.57	53.06	30.61	44.90
	Block 3	91.98 (1.22↓)	57.14	77.55	10.20	28.57	14.29	24.49	30.61	36.73

Table 8. **Performance of MIAs under defenses, with FFHQ as the image prior.** The private dataset is VGGFace2, red refers to MobileNetV2, and blue refers to Swin Transformer.

D. The performance of long-tailed surrogate training on the private dataset

We evaluate surrogate model performance using VGGFace2 as the private dataset with varying sample sizes, as shown in Tab. 9. When the sample size is set to $2.5K$, surrogate models trained on the private dataset (M_s^{priv}) perform worse than those trained on the public dataset (M_s^{pub}). As the sample size increases, M_s^{priv} gradually approach or surpass M_s^{pub}. This phenomenon is explained in Fig. 1(a), where the public dataset results in lower Top-1 confidence scores, indicating flat model outputs, whereas the private dataset yields higher Top-1 confidence scores. We believe flat outputs are more beneficial for surrogate models in shaping decision boundaries with limited training data, while higher confidence provides clearer category information when more data is available. The general performance trend of long-tailed surrogate training is shown in Fig. 1(b).

D_{priv}	VGGFace2,InceptionV1				VGGFace2,ResNet50			
D	CelebA	FFHQ	VGGFace2	VGGFace2	CelebA	FFHQ	VGGFace2	VGGFace2
InceptionV1*	21.84/39.29	36.54/58.90	11.74/20.36	10.83/15.36	8.91/19.40	20.45/38.98	8.23/16.42	8.19/16.99
EfficientNetB0	9.80/20.74	17.46/33.22	7.66/13.65	5.37/9.36	3.08/7.76	7.01/16.55	3.05/6.82	2.79/5.69
InceptionV1*	20.63/37.44	40.17/61.00	21.49/34.39	18.61/26.48	9.16/19.67	23.78/42.76	18.90/27.43	17.85/25.95
EfficientNetB0	15.45/29.81	24.06/42.86	13.86/22.55	9.18/15.43	4.75/11.59	11.57/25.07	5.09/9.84	4.98/9.54
InceptionV1*	24.87/42.08	45.12/64.68	39.64/56.82	25.84/38.18	14.27/28.28	32.29/52.51	33.24/45.67	27.98/40.03
EfficientNetB0	20.91/37.94	31.54/51.59	27.14/41.88	20.04/30.92	7.13/16.81	17.51/34.26	12.02/20.36	10.26/18.15

Table 9. We set the sample sizes to $2.5K$, $5K$, and $10K$. VGGFace2 refers to training surrogate models on outputs of the target model using private dataset, while VGGFace2 refers to using the hard labels corresponding to the private dataset.

E. More results about Long-tailed Learning

A possible strategy to alleviate the long-tail issue is to use auxiliary priors with better diversity (as discussed in Section 3.2.). From the attacker's perspective, our goal is not to rely on better priors or larger sample sizes but to extract more information from extreme long-tail distributions to improve surrogate models. This alleviates the performance degradation of surrogate models caused by the long-tail issue, but does not resolve the long-tail issue itself. We further show the boost that long-tailed surrogate training brings to classification, shown in Fig. 2. The overall performance improvement can be clearly observed.

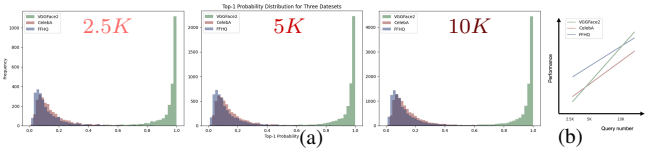


Figure 1. **Distribution of Top-1 confidence scores.** The target model is set to VGGFace2,ResNet50.

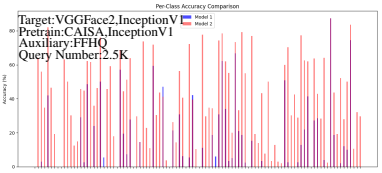


Figure 2. Model 1 is from *Base*, Model 2 is from long-tailed surrogate training. The figure shows the model's performance on the first 100 IDs.

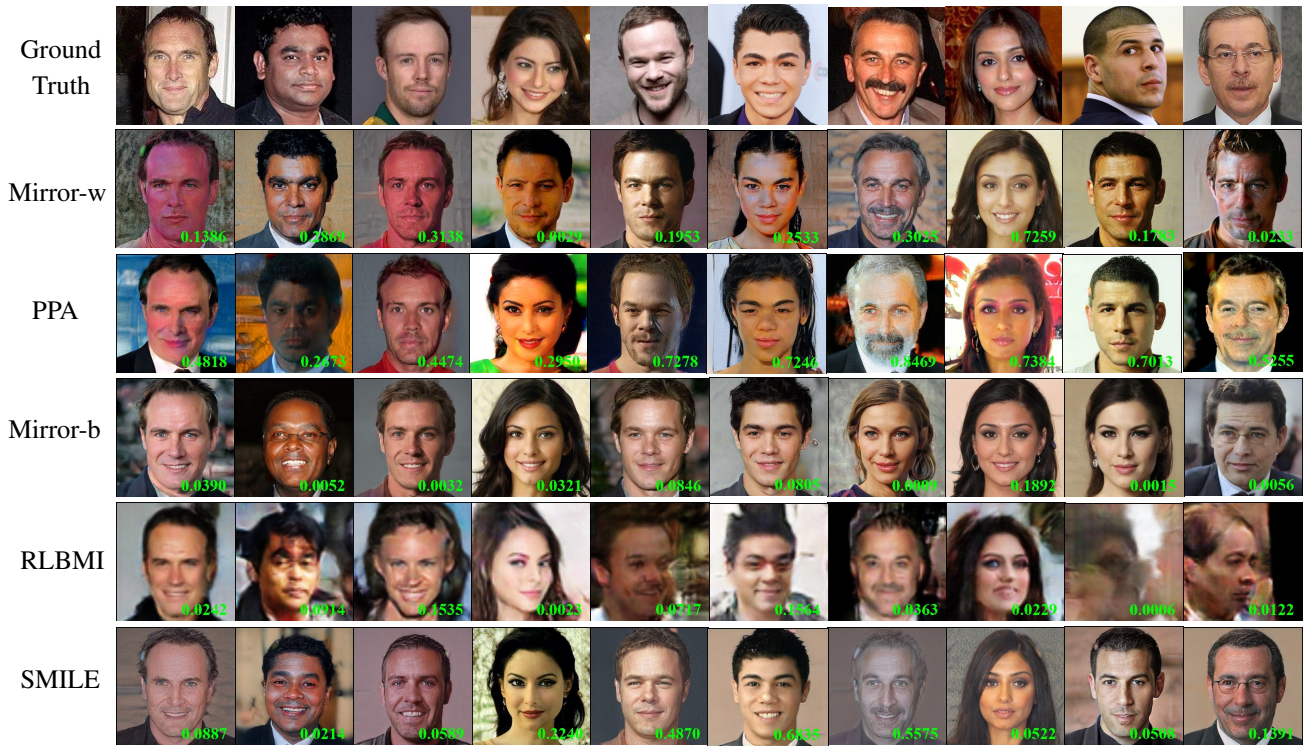
F. More experiments under CASIA pre-trained models

In MIAs, IID refers to splitting a dataset into private and auxiliary parts (both highly aligned). We chose VGGFace2 pre-trained models for their diverse architectures, which help validate our method's robustness. The data distributions of VGGFace and VGGFace2 are relatively close, which may cause concern. Therefore, we add experiments under CASIA pre-trained models (Tab. 10). We believe that the similar attack performance is due to the alignment and distribution differences between the target model/pre-trained model and the synthetic data, which weakens the impact of the pre-trained model's training data.

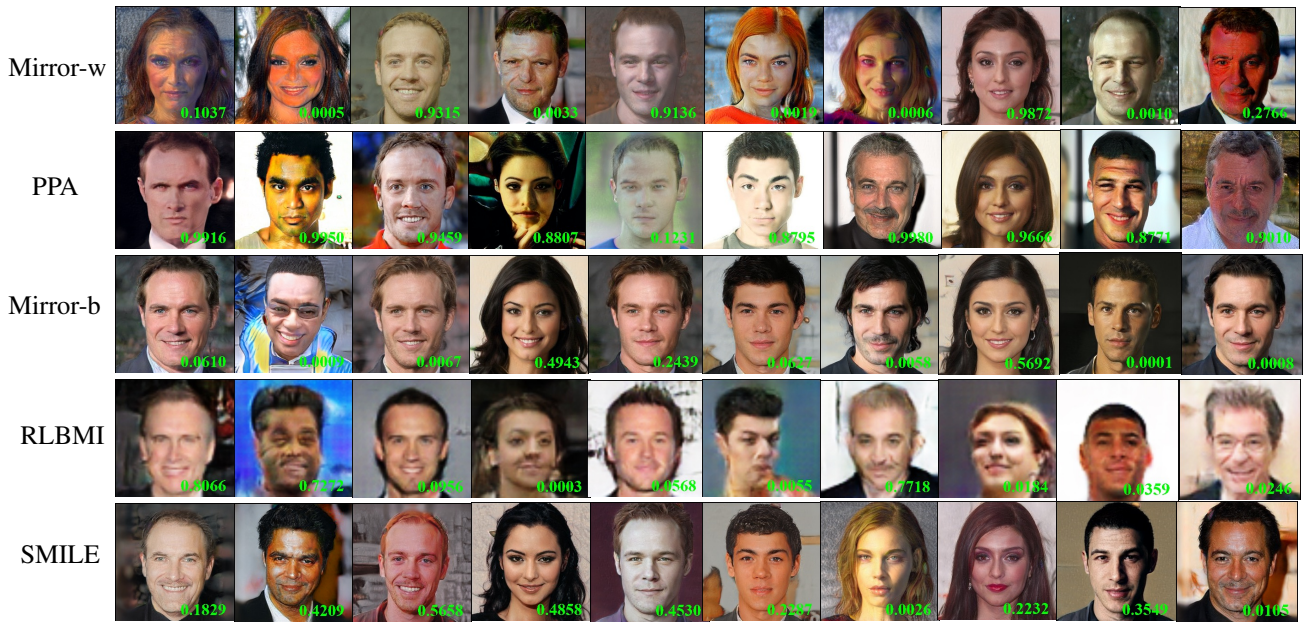
D_{priv}	VGGFace							
D_{pub}	CelebA				FFHQ			
M_t	VGG16		VGG16BN		VGG16		VGG16BN	
Method	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Average	71.28	85.91	68.56	79.79	77.89	90.02	66.3	84.23
EfficientNetB0	69.39 ± 3.33	82.31 ± 1.92	70.75 ± 0.96	76.87 ± 0.96	80.95 ± 3.46	93.20 ± 0.96	68.03 ± 2.54	87.07 ± 1.92
InceptionV1*	65.99 ± 2.54	83.67 ± 1.66	64.62 ± 2.54	82.31 ± 2.54	78.91 ± 3.46	91.16 ± 0.96	66.66 ± 3.85	80.95 ± 2.54

Table 10. **Results under CASIA pre-trained models. Average** is the mean result across different architectures of VGGFace2 pre-trained models.

G. More qualitative results



VGGFace2 / ResNet50 / CelebA
(D_{priv} / M_t / Image prior)



VGGFace2 / InceptionV1 / CelebA

VGGFace2 / ResNet50 / FFHQ

VGGFace2 / InceptionV1 / FFHQ

Figure 3. **More qualitative results.** The surrogate model used by *SMILE* is InceptionV1 pre-trained on CASIA.