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.

$\mathcal{D}_{priv}$	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.

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.

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.

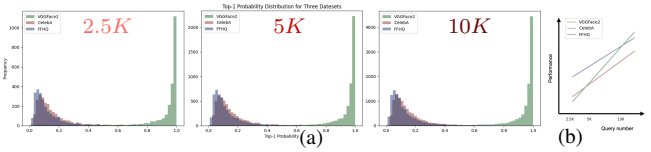


Figure 1. **Distribution of Top-1 confidence scores.** The target model is set to VGGFace2,ResNet50.

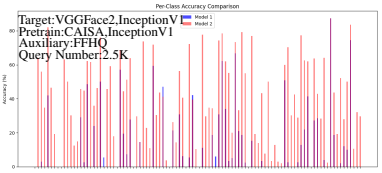


Figure 2. Model 1 is from *Base*, Model 2 is from long-tailed surrogate training. The figure shows the model's performance on the first 100 IDs.

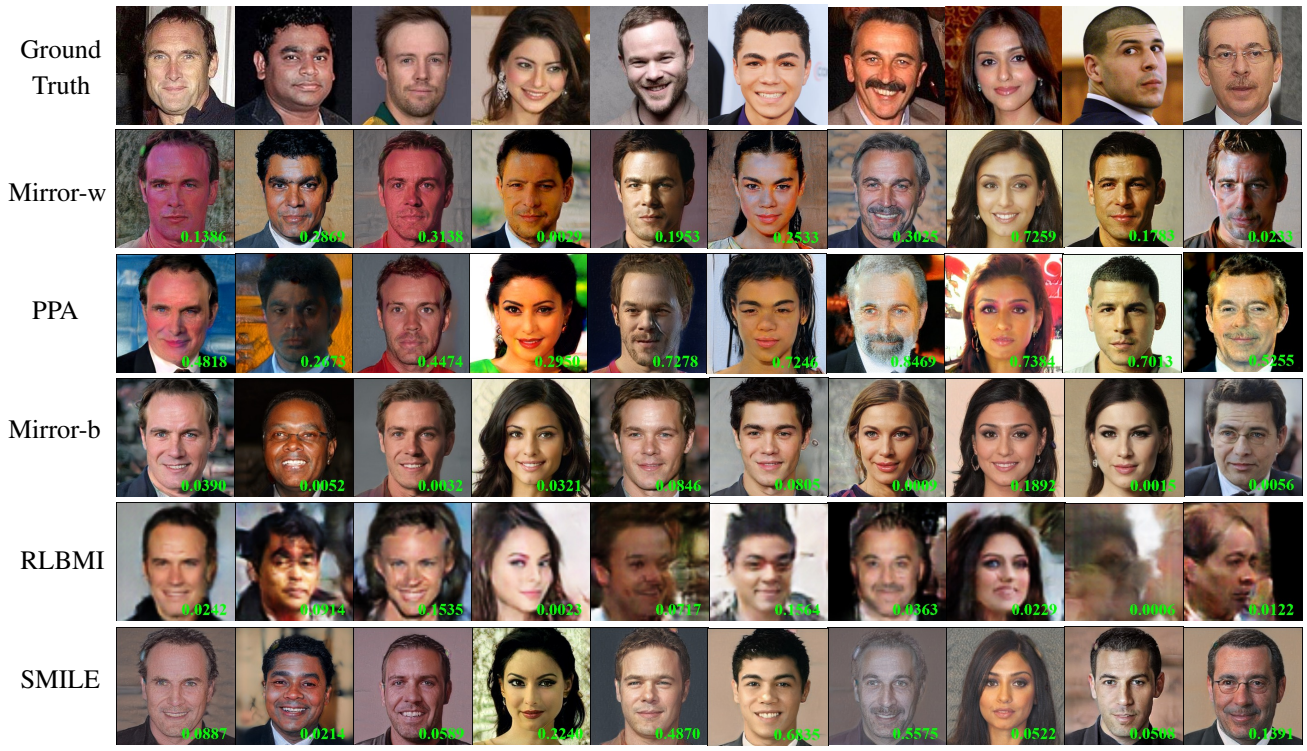
F. More experiments under CASIA pre-trained models

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.

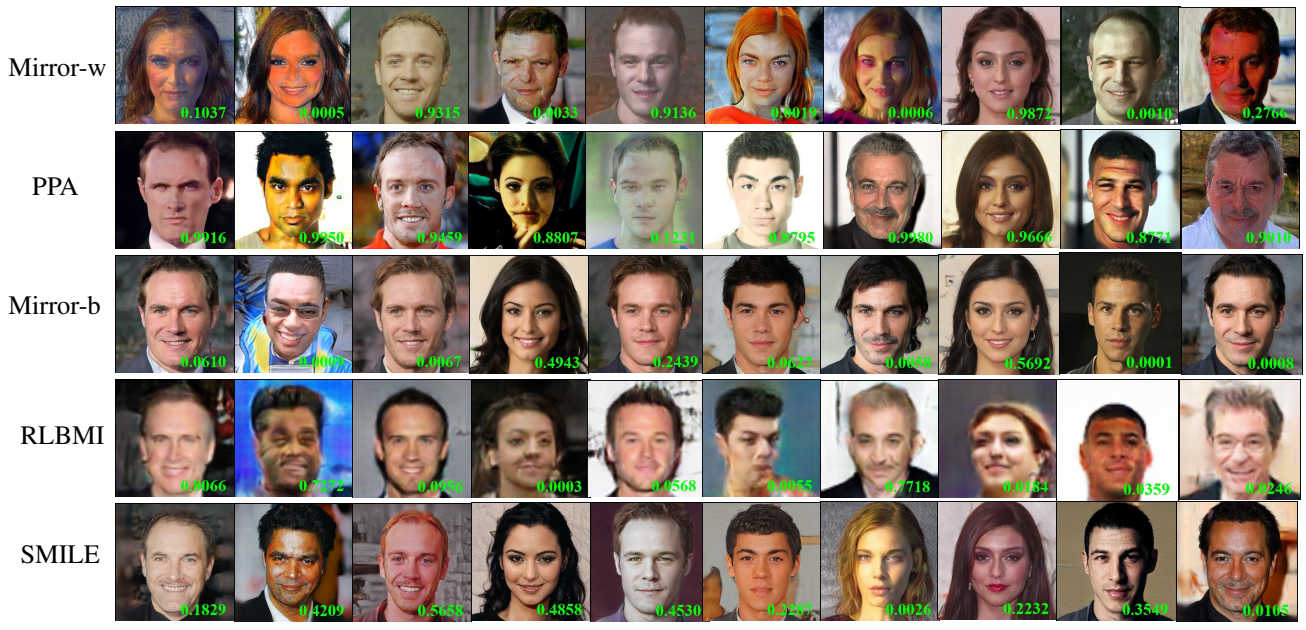
D_{priv}	VGGFace							
D_{pub}	CelebA				FFHQ			
M_t	VGG16		VGG16BN		VGG16		VGG16BN	
Method	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Average	71.28	85.91	68.56	79.79	77.89	90.02	66.3	84.23
EfficientNetB0	69.39 ± 3.33	82.31 ± 1.92	70.75 ± 0.96	76.87 ± 0.96	80.95 ± 3.46	93.20 ± 0.96	68.03 ± 2.54	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.