Re-thinking Model Inversion Attacks Against Deep Neural Networks Supplementary Materials

Ngoc-Bao Nguyen[*] Keshigeyan Chandrasegaran[*] Milad Abdollahzadeh Ngai-Man Cheung[†]
Singapore University of Technology and Design (SUTD)
thibaongoc_nguyen@mymail.sutd.edu.sg, {keshigeyan, milad_abdollahzadeh, ngaiman_cheung}@sutd.edu.sg

Overview

In this supplementary material, we provide additional experiments, analysis, ablation study, and reproducibility details to support our findings. We provide Pytorch code, demo and pre-trained models (target models/ evaluation models/ augmented models) at: https://ngoc-nguyen-0.github.io/re-thinking_model_inversion_attacks/.

Contents

A. Additional experimental results

In this section, we provide additional experimental results that are not included in the main paper. More specifically, first, we evaluate the effect of the proposed method on improving SOTA approaches in new tasks including image classification and digit classification. Then, we use alternative metrics for evaluating SOTA MI approaches with and without proposed improvements on identity loss L_{id}. The additional experimental results in this section further support effectiveness of the proposed approach on improving MI attacks.

A.1. Experimental results on CIFAR-10 and MNIST

In Sec. 4.2 of the main paper, we mostly focus on the face recognition task (on the CelebA dataset) and show that the proposed method significantly improves SOTA approaches by increasing **Attack Acc** (inference accuracy on reconstructed samples by an evaluation model; see Sec. 4.1. of the main paper) and decreasing **KNN Dist** (distance between the reconstructed samples of a specific class/id and corresponding data in the private dataset $\mathcal{D}_{priv}$; see Sec. 4.1).

In this section, we provide results for other tasks. More specifically, as mentioned in Sec. 4.1, for GMI [20], and KEDMI [1], following their own setup, we use digit classification task MNIST dataset, and object classification task on the CIFAR-10 dataset. For each task, Table 1 tabulates the performance of the SOTA approach together with three variants of our proposed approach:

1. + LOM (Ours): We replace existing identity loss, L_{id} with our improved identity loss L_{id}^{logit} (Sec. 3.1).
2. + MA (Ours): We replace existing identity loss, L_{id} with our proposed L_{id}^{aug} (Sec. 3.2).
3. + LOMMA (Ours): We combine both L_{id}^{logit} and L_{id}^{aug} for model inversion.

[*] Equal Contribution [†] Corresponding Author

Table 1. We report top 1 accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance for two experiment setups. Following exact experiment setups in [1]. For CIFAR-10 experiments, $\mathcal{D}_{priv}$ = CIFAR-10, $\mathcal{D}_{pub}$ = CIFAR-10, M_t = VGG16, evaluation model = ResNet-18. For MNIST experiments, $\mathcal{D}_{priv}$ = MNIST, $\mathcal{D}_{pub}$ = MNIST, M_t = CNN(Conv3), evaluation model = CNN(Conv5). The best results are in **bold**.

Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
CIFAR-10/CIFAR-10/VGG16			
KEDMI	95.2 ± 7.96	-	78.24
+ LOM (Ours)	100 ± 0	4.80	**52.12**
+ MA (Ours)	100 ± 0	4.80	53.17
+ LOMMA (Ours)	**100 ± 0**	**4.80**	63.41
GMI	43.20 ± 19.80	-	96.11
+ LOM (Ours)	80.80 ± 14.65	37.60	**70.47**
+ MA (Ours)	80.00 ± 18.01	36.80	93.46
+ LOMMA (Ours)	**95.20 ± 7.96**	**52.00**	80.30
MNIST/MNIST/CNN(Conv3)			
KEDMI	46.40 ± 14.65	-	120.99
+ LOM (Ours)	55.20 ± 8.94	8.80	100.18
+ MA (Ours)	75.20 ± 6.57	28.80	72.38
+ LOMMA (Ours)	**100.00 ± 0.00**	**53.60**	**58.81**
GMI	8.00 ± 1.52	-	126.61
+ LOM (Ours)	15.20 ± 15.12	7.20	161.90
+ MA (Ours)	66.40 ± 19.86	58.40	**78.38**
+ LOMMA (Ours)	**80.80 ± 17.38**	**72.80**	83.56

Table 2. We follow exact the experiment setup of [17] for the VMI experiments. Specifically, we use DCGAN and Flow model to learn the distribution of $\mathbf{z}$.

Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
MNIST/EMNIST/ResNet-10			
VMI	94.60 ± 0.13	-	68.53
+ LOM (Ours)	98.60 ± 0.09	4.00	88.13
+ MA (Ours)	98.98 ± 0.02	4.38	58.81
+ LOMMA (Ours)	**100.00** ± 0.00	5.40	**52.62**

As one can see, on average each of the proposed solutions drastically improves the SOTA approaches, and combining these two solutions works even better.

Additionally, as mentioned in Sec. 4.1, for VMI [17], following the setup in [17], we evaluate its performance for digit classification on MNIST, and improvement brought by the proposed method. Note that for a fair comparison, following VMI implementation in [17], in this experiment we use EMNIST [4] as public dataset $\mathcal{D}_{pub}$ to acquire prior knowledge. Results are shown in Table 2 for three variants of our proposed method, which indicates better performance in terms of both attack accuracy (reaching 100% attack accuracy) and decreasing KNN Distance.

Table 3. We report the results for KEDMI and GMI for IR152, face.evoLve and VGG16 target model. Following exact experiment setups in [1], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, evaluation model = face.evoLve. We report top-5 accuracies, the improvement compared to the SOTA MI (Imp.), and FID scores.

Method	Top-5 Attack Acc ↑	Imp. ↑	FID ↓
CelebA/CelebA/IR152			
KEDMI	98.00 ± 1.96	-	**28.06**
+ LOM (Ours)	**98.67** ± 0.00	**0.67**	39.03
+ MA (Ours)	98.33 ± 1.19	0.33	28.38
+ LOMMA (Ours)	**98.67** ± 0.37	**0.67**	36.78
GMI	55.67 ± 7.14	-	57.11
+ LOM (Ours)	93.00 ± 3.41	37.33	48.87
+ MA (Ours)	89.00 ± 4.10	33.33	45.24
+ LOMMA (Ours)	**97.67 ± 2.41**	**42.00**	**45.02**
CelebA/CelebA/face.evoLve			
KEDMI	97.33 ± 1.73	-	**31.26**
+ LOM (Ours)	**99.33** ± 0.18	**2.00**	42.45
+ MA (Ours)	98.00 ± 0.94	0.67	32.08
+ LOMMA (Ours)	**99.33** ± 0.33	**2.00**	38.69
GMI	45.33 ± 8.05	-	59.76
+ LOM (Ours)	84.33 ± 4.49	39.00	44.27
+ MA (Ours)	92.00 ± 2.25	46.67	51.15
+ LOMMA (Ours)	**93.67 ± 2.42**	**48.33**	**44.07**
CelebA/CelebA/VGG16			
KEDMI	93.33 ± 3.36	-	25.46
+ LOM (Ours)	**99.00 ± 0.18**	**5.67**	34.45
+ MA (Ours)	95.33 ± 1.60	2.00	**24.65**
+ LOMMA (Ours)	98.00 ± 0.61	4.67	33.91
GMI	40.33 ± 4.74	-	58.03
+ LOM (Ours)	89.33 ± 2.73	49.00	46.40
+ MA (Ours)	81.33 ± 5.88	41.00	44.90
+ LOMMA (Ours)	**95.67 ± 2.16**	**55.34**	**43.21**

A.2. Experimental Results with Additional Metrics

As mentioned in Sec. 4.1 of the main paper, Attack Acc and KNN Dist are common metrics used in literature to evaluate the MI attacks. In this section, we include results on two additional metrics namely: Top-5 Attack Acc and FID [9]. Results in Table 3, Table 4, and Table 5 show that the proposed method achieves better performance in terms of Top-5 Attack Acc, and FID value.

B. Ablation Study

B.1. Different number of augmented models M_{aug}

In Sec 3.2, we propose a model augmentation idea with augmented models M_{aug}. Here, we experiment using a different number of networks for M_{aug}. Table 6 show that increasing the number of the augmented models will improve attack accuracy. We use 3 augmented models in our main result as this configuration achieves a good tradeoff in accuracy and computation.

Table 4. We report the results for VMI . Following exact experiment setups in [17], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, M_t = ResNet-34, evaluation model = IR-SE50. We report top-5 accuracies, the improvement compared to the SOTA MI (Imp.), and FID scores.

Method	Top-5 Attack Acc ↑	Imp. ↑	FID ↓
CelebA/CelebA/ResNet-34			
VMI	82.32 ± 0.21	-	**16.82**
+ LOM (Ours)	86.56 ± 0.27	4.24	25.42
+ MA (Ours)	86.16 ± 0.19	3.84	17.60
+ LOMMA (Ours)	**91.02 ± 0.22**	**8.70**	23.56

Table 5. We report the results for KEDMI and GMI for IR152, face.evoLve and VGG16 target model. Following exact experiment setups in [1], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = FFHQ, evaluation model = face.evoLve. We report top-5 accuracies, the improvement compared to the SOTA MI (Imp.), and FID scores.

Method	Top-5 Attack Acc ↑	Imp. ↑	FID ↓
CelebA/FFHQ/IR152			
KEDMI	85.33 ± 4.01	-	41.71
+ LOM (Ours)	88.67 ± 1.18	3.33	50.84
+ MA (Ours)	87.67 ± 2.28	2.33	**39.88**
+ LOMMA (Ours)	**92.00 ± 0.57**	**6.60**	45.67
GMI	36.33 ± 3.98	-	47.72
+ LOM (Ours)	80.33 ± 4.21	44.00	40.18
+ MA (Ours)	84.00 ± 5.35	47.67	**35.41**
+ LOMMA (Ours)	**90.33 ± 3.16**	**54.00**	37.58
CelebA/FFHQ/face.evoLve			
KEDMI	80.67 ± 2.83	-	38.09
+ LOM (Ours)	91.33 ± 0.47	10.67	47.30
+ MA (Ours)	88.67 ± 2.44	8.00	**35.94**
+ LOMMA (Ours)	**94.00 ± 0.68**	**13.33**	47.51
GMI	33.33 ± 6.18	-	52.84
+ LOM (Ours)	74.67 ± 4.78	41.33	44.01
+ MA (Ours)	72.00 ± 4.64	38.67	**35.58**
+ LOMMA (Ours)	**89.00 ± 2.73**	**55.67**	40.03
CelebA/FFHQ/VGG16			
KEDMI	74.00 ± 4.05	-	36.18
+ LOM (Ours)	81.67 ± 1.19	7.67	43.76
+ MA (Ours)	80.33 ± 3.27	6.33	**35.02**
+ LOMMA (Ours)	**85.33 ± 1.98**	**11.33**	40.26
GMI	25.67 ± 5.13	-	53.17
+ LOM (Ours)	70.67 ± 3.92	45.00	42.60
+ MA (Ours)	62.33 ± 5.36	36.67	36.04
+ LOMMA (Ours)	**86.33 ± 5.17**	**60.67**	**35.59**

B.2. Different network architectures for M_{aug}

In this section, we provide additional results by using different structures for augmenting the target model using M_{aug} in the MI process. Note that the architecture of all these models is different from the one used for target model M_t.

More specifically, we use three different combinations for M_{aug}, each of which contains three models: (i) {EfficientNet-B0, EfficientNet-B1, EfficientNet-B2}, and (ii) {DenseNet121, DenseNet161, DenseNet169}, and (iii) {EfficientNet-B0, DenseNet121, MobileNetV3}. Results in Table 7 shows that $+MA$ (Ours) consistently improves the attack accuracy and KNN distance with different network architectures.

B.3. The effect of different sizes of public dataset

We conduct additional experiments using different sizes of $\mathcal{D}_{pub}$ (10%, 50%) to emulate the different quality of prior information. The results for KEDMI [1] are shown in Table 8. The key observations are:

- Baseline attack accuracies are poorer under limited $\mathcal{D}_{pub}$, *i.e.* $\mathcal{D}_{pub}$ = 10%.
- Our proposed method can outperform existing SOTA under varying degrees of prior information although the improvement obtained by KD is marginal under $\mathcal{D}_{pub}$ = 10%.

C. Additional analysis and details on experimental setups

C.1. Details on combining L_{id}^{logit} and L_{id}^{aug}

We provide details of combining L_{id}^{logit} and L_{id}^{aug}. We substitute L_{id}^{logit} (Eqn. 3 of main paper) into L_{id}^{aug} (Eqn. 4 of main paper) for an inversion targeting class k of the target model M_t, using augmented model $M_{aug}^{(i)}$. In particular, starting from Eqn. 4 of the main paper:

$$\begin{aligned}
L_{id}^{aug}(\mathbf{x}; y) \quad &= \quad \gamma_t \cdot L_{id}(\mathbf{x}; y, M_t) \\
&\quad\quad +\gamma_{aug} \cdot \sum_{i=1}^{N_{aug}} L_{id}(\mathbf{x}; y, M_{aug}^{(i)}) \\
&= \quad \gamma_t \cdot L_{id}^{logit}(\mathbf{x}; y, M_t) \\
&\quad\quad +\gamma_{aug} \cdot \sum_{i=1}^{N_{aug}} L_{id}^{logit}(\mathbf{x}; y, M_{aug}^{(i)}) \\
&= \quad \gamma_t \cdot (-\log \mathbf{p}_t^T \mathbf{w}_{t,k} + \lambda ||\mathbf{p}_t - \mathbf{p}_{reg}||_2^2) \\
&\quad\quad +\gamma_{aug} \cdot \sum_{i=1}^{N_{aug}} (-\log{(\mathbf{p}_{aug}^{(i)})^T}(\mathbf{w}_{aug,k}^{(i)}) \\
&\quad\quad +\lambda ||\mathbf{p}_{aug}^{(i)} - \mathbf{p}_{reg}||_2^2) \\
&\approx \quad \gamma_t \cdot (-\log \mathbf{p}_t^T \mathbf{w}_{t,k}) \\
&\quad\quad +\gamma_{aug} \cdot \sum_{i=1}^{N_{aug}} (-\log{(\mathbf{p}_{aug}^{(i)})^T}(\mathbf{w}_{aug,k}^{(i)})) \\
&\quad\quad +\lambda' ||\mathbf{p}_t - \mathbf{p}_{reg}||_2^2 \qquad (1)
\end{aligned}$$

Table 6. We report top-1 attack accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance for using different numbers N_{aug} of network M_{aug}. Following exact experiment setups in [1], here method = KEDMI, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, M_t = IR152, evaluation model = face.evoLve. We select M_{aug} from the set of 4 networks including EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3. The number of network M_{aug} increases from 0 (Baseline KEDMI) to 4. It shows that using more M_{aug} improves the attack accuracy and KNN distance.

Method	N_{aug}	M_{aug}	Attack Acc ↑	Imp. ↑	KNN dist ↓
		CelebA/CelebA/IR152			
KEDMI	-	-	80.53 ± 3.86	-	1247.28
+ MA	1	EfficientNet-B0	81.20 ± 3.75	0.67	1234.16
+ MA	2	EfficientNet-B0, EfficientNet-B1	84.47 ± 2.99	3.94	1223.56
+ MA	3	EfficientNet-B0, EfficientNet-B1, EfficientNet-B2	84.73 ± 3.76	4.20	1220.23
+ MA	4	EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3	**85.87 ± 2.63**	**5.34**	**1217.15**

Table 7. We report top-1 attack accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance for different structures of network M_{aug}. Following exact experiment setups in [1], here method = KEDMI, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, M_t = IR152, evaluation model = face.evoLve. We select different network architectures for our experiment. Specifically, we use Ours-1 = {EfficientNet-B0 [16], EfficientNet-B1 [16], EfficientNet-B2 [16]}, Ours-2 = {DenseNet121 [11], DenseNet161 [11], DenseNet169 [11]}, Ours-3 = {EfficientNet-B0, DenseNet121 [11], MobileNetV3-large [10]}. It shows that using different network architectures M_{aug} consistently improves the attack accuracy and KNN distance.

Method	M_{aug}	Attack Acc ↑	Imp. ↑	KNN dist ↓
	CelebA/CelebA/IR152			
KEDMI	-	80.53 ± 3.86	-	1247.28
+ MA (Ours-1)	EfficientNet-B0, EfficientNet-B1, EfficientNet-B2	**84.73 ± 3.76**	**4.20**	**1220.23**
+ MA (Ours-2)	DenseNet121, DenseNet161, DenseNet169	**89.07 ± 3.32**	**8.54**	**1211.73**
+ MA (Ours-3)	EfficientNet-B0, DenseNet121, MobileNetV3-large	**86.53 ± 1.98**	**6.00**	**1204.94**

Here, $\mathbf{p}_t$, $\mathbf{w}_{t,k}$ are penultimate layer activation and last layer weight for the target model M_t; $\mathbf{p}_{aug}^{(i)}$, $\mathbf{w}_{aug,k}^{(i)}$ are penultimate layer activation and last layer weight for the augmented model $M_{aug}^{(i)}$. Note that one regularization is sufficient as shown in the last step. Eqn. 1 above is used in Eqn. 1 of the main paper in the inversion step using the proposed method.

C.2. Details on improving KEDMI baseline

We apply a simple technique that is introduced by GMI [20] to get better results for KEDMI [1]. Specifically, after model inversion, and sampling $\mathbf{z}$ from the learned distribution, we clip all elements of $\mathbf{z}$ into $[-1, 1]$, which is shown to be beneficial in [20]. In Table 9, we observe that clipping $\mathbf{z}$ help to boost the attack accuracy of KEDMI and the reconstructed images are more similar to the private dataset as KNN distances are reduced. Therefore, for all the experiments with KEDMI in the main paper and Supp, we clip $\mathbf{z}$ to get better results and we compare with this better version of KEDMI.

C.3. Additional details on computing $\mathbf{p}_{reg}$

In Sec. 3.1, we propose an improved formulation for identity loss L_{id}^{logit} which includes a regularization term $||\mathbf{p} - \mathbf{p}_{reg}||_2^2$ to prevent unbound growth of norm during optimization. Here we provide additional details on computing $\mathbf{p}_{reg}$.

Given that the attacker has no access to private training data, we estimate $\mathbf{p}_{reg}$ by a simple method using *public* data. We firstly construct the set of penultimate layer features of public data using the target model and estimate the mean μ_{pen} and variance σ_{pen}^2:

$$\mu_{pen} = \frac{1}{N}\sum_{i=1}^{N} M^{pen}(\mathbf{x}_i) \tag{2}$$

$$\sigma_{pen}^2 = \frac{1}{N}\sum_{i=1}^{N}(M^{pen}(\mathbf{x}_i) - \mu_{pen})^2 \tag{3}$$

where $\mathbf{x}_i$ is a sample from public dataset $\mathcal{D}_{pub}$, and $M^{pen}()$ operator returns the penultimate layer representations of the target model M_t for a given input $\mathbf{x}$. We analyze two ways to estimate $\mathbf{p}_{reg}$ as follow:

- Fixed $\mathbf{p}_{reg}$ where $\mathbf{p}_{reg} = \mu_{pen}$.
- $\mathbf{p}_{reg}$ is sampled using the prior distribution $\mathcal{N}(\mu_{pen}, \sigma_{pen})$.

Empirically, we use $N = 5,000$ images from the public dataset $\mathcal{D}_{pub}$ to estimate μ_{pen} and σ_{pen}. The results show that using $\mathbf{p}_{reg}$ which is sampled from $\mathcal{N}(\mu_{pen}, \sigma_{pen})$ gives

Table 8. Sensitivity of the proposed method to prior information, $\mathcal{D}_{pub}$: We use $\mathcal{D}_{priv}/\mathcal{D}_{pub}$ = CelebA, M_t = face.evoLve, evaluation = face.evoLve and KEDMI [1]. We report top 1 MI attack accuracy and KNN distance using 10%, 50% and 100% of D_{pub}. As GAN is trained on D_{pub}, it affects the baseline KEDMI and our proposed method. The results show that + LOM and + MA consistently improve upon the baseline.

	$\mathcal{D}_{pub}$ **= 10%**		$\mathcal{D}_{pub}$ **= 50%**		$\mathcal{D}_{pub}$ **= 100%**	
	Attack Acc ↑	**KNN Dist ↓**	**Attack Acc ↑**	**KNN Dist ↓**	**Attack Acc ↑**	**KNN Dist ↓**
KEDMI	58.33 ± 5.25	1450.06	79.07 ± 3.76	1265.37	81.40 ± 3.25	1248.32
+ LOM (Ours)	67.27 ± 1.83	1395.38	89.27 ± 0.96	1202.45	92.53 ± 1.51	1183.76
+ MA (Ours)	61.80 ± 3.03	1421.83	82.20 ± 2.77	1244.21	85.07 ± 2.71	1222.02
+ LOMMA (Ours)	**74.40 ± 2.21**	**1328.79**	**89.67 ± 0.76**	**1170.37**	**93.20 ± 0.85**	**1154.32**

Table 9. We apply a simple technique that is introduced by GMI [20] to get better baseline results for KEDMI [1]. We report the results for KEDMI with and without **z** clipping for IR152, face.evoLve, and VGG16 target model. Following exact experiment setups in [1], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, evaluation model = face.evoLve. We report top-1 attack accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance. The improvement using **z** clipping is clear.

Method	Attack Acc ↑	Imp. ↑	KNN dist ↓
CelebA/CelebA/IR152			
KEDMI w/o **z** clipping	78.53 ± 3.45	-	1270.87
KEDMI with **z** clipping	**80.53 ± 3.86**	**2.00**	**1247.28**
CelebA/CelebA/face.evoLve			
KEDMI w/o **z** clipping	78.00 ± 4.09	-	1290.62
KEDMI with **z** clipping	**81.40 ± 3.25**	**3.40**	**1248.32**
CelebA/CelebA/VGG16			
KEDMI w/o **z** clipping	67.93 ± 4.24	-	1345.03
KEDMI with **z** clipping	**74.00 ± 3.10**	**6.07**	**1289.88**

Table 10. We report the results for KEDMI using a fixed $\mathbf{p}_{reg}$ or sampling from a distribution approximated for $\mathbf{p}_{reg}$. We use three different target models: IR152, face.evoLve, and VGG16. Following exact experiment setups in [1], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, evaluation model = face.evoLve. We report top-1 attack accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance.

Method	Attack Acc ↑	KNN dist ↓
CelebA/CelebA/IR152		
+ LOM (Fixed p_{reg})	92.27 ± 1.37	**1155.92**
+ LOM (Ours)	**92.47 ± 1.41**	1168.55
CelebA/CelebA/face.evoLve		
+ LOM (Fixed p_{reg})	90.40 ± 1.68	1257.95
+ LOM (Ours)	**92.53 ± 1.51**	**1183.76**
CelebA/CelebA/VGG16		
+ LOM (Fixed p_{reg})	85.60 ± 1.79	1259.60
+ LOM (Ours)	**89.07 ± 1.46**	**1218.46**

better performance than using fixed $\mathbf{p}_{reg} = \mu_{pen}$ (see Table 10). Therefore, all the results reported in the main paper use the $\mathbf{p}_{reg} \sim \mathcal{N}(\mu_{pen}, \sigma_{pen})$. We remark again that $\mathbf{p}_{reg}$ is estimated from *public* dataset.

C.4. Details on regularization parameter λ

In Sec 3.1 of the main paper, the regularization term $||\mathbf{p}-\mathbf{p}_{reg}||_2^2$ includes a parameter λ which controls the effect of this term. In this section, we evaluate the effect of this parameter by examining different values of λ on model inversion performance. Results in Table 11 show that attack accuracy is improved over SOTA KEDMI with our proposed logit loss even without the regularization term ($\lambda = 0$). However, we get better results if the regularization is added *e.g.* $\lambda = 1.0$. Due to its better performance, we use $\lambda = 1.0$ in all experiments with the proposed method.

C.5. Computational overhead

In order to investigate the computational overhead introduced by our proposed method, in this section, we report the running time for reconstructing images of 300 identities on CelebA/CelebA/IR152 setup for KEDMI and GMI, and 100 identities on CelebA/CelebA/ResNet-34 for VMI. All the experiments of KEDMI and GMI are performed on an NVIDIA GeForce RTX 3090 GPU, and the experiments of VMI are performed on an NVIDIA RTX A5000 GPU. The results in Table 12 show that + LOM does not affect the training time compared to the baseline. However, + MA adds some computational overhead as it uses additional net-

Table 11. We report the results for KEDMI with different λ values using IR152 as target model. Following exact experimental setups in [1], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, evaluation model = face.evoLve. We report top-1 attack accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance.

Method	λ	Attack Acc ↑	Imp. ↑	KNN dist ↓
CelebA/CelebA/IR152				
KEDMI	-	80.53 ± 3.86	-	1247.28
+ LOM	0	90.33 ± 1.64	9.80	1198.39
+ LOM	0.5	89.53 ± 1.21	9.00	1175.35
+ LOM	1.0	**92.47 ± 1.41**	**11.94**	1168.55
+ LOM	2.0	91.87 ± 1.09	11.34	1125.54
+ LOM	10.0	85.80 ± 1.24	5.27	**1110.80**

works M_{aug} during the inversion.

Table 12. Computational complexity of different algorithms in terms of average running time (GPU hours) using single GPU. We use KEDMI, GMI, and VMI approaches as the baseline. We have also included the running time Ratio when compared to the corresponding baseline.

Method	**RunTime (hrs)**	**Ratio**
KEDMI	0.35	1.00
+ LOM (Ours)	0.35	1.00
+ MA (Ours)	0.60	1.71
+ LOMMA (Ours)	0.60	1.72
GMI	1.61	1.00
+ LOM (Ours)	1.61	1.00
+ MA (Ours)	2.83	1.76
+ LOMMA (Ours)	2.85	1.77
VMI	364.67	1.00
+ LOM (Ours)	368.24	1.01
+ MA (Ours)	368.69	1.01
+ LOMMA (Ours)	379.41	1.04

C.6. Hyperparameters

In the experiments of GMI and KEDMI, we do the inversion using SGD optimizer with the learning rate $lr = 0.02$ in 2400 iterations which are used from the released code of KEDMI [1]. We set $\gamma_t = \gamma_{aug} = 100/(N_{aug} + 1)$ and $\lambda = 100$, where N_{aug} is the number of models used for augmented model M_{aug}. We estimate $\mathbf{p}_{reg}$ for each classifier by using $N = 5,000$ images from the public dataset $\mathcal{D}_{pub}$. In the experiments of VMI, we use 20 epochs (equal to 3120 iterations) to learn the distribution of each identity.

C.7. Dataset

Experiments of KEDMI and GMI. We follow exact experimental setups in [1]. For the CelebA task, we use the dataset divided by [1] for all of the experiments. In particular, the private dataset has 30,027 images of 1000 identities and the public dataset has 30000 images that are non-overlapping identities with the private dataset. In the experiments in Table 5 (main paper), we use FFHQ [13] as the public dataset to train GAN and distill knowledge to augmented models. For MNIST and CIFAR-10 tasks, the private dataset contains images with labels from 0 to 4 and the public dataset includes the rest of the dataset with labels from 5 to 9.

Experiments of VMI. We follow exact experimental setup in [17]. We use the CelebA dataset and MNIST dataset for VMI experiments. For CelebA, we follow [17] to divide the dataset into two parts. The first part contains images of 1000 most frequent identities which uses as private dataset. The rest of dataset is used as public dataset. For the experiments on MNIST dataset, we use EMNIST [4] as public dataset to train GAN and M_{aug}.

D. Additional Visualizations

D.1. Additional Results for GMI

Similar to results reported for KEDMI (Figure 4, main paper), in this section, we show results for GMI [20] under IR152 target classifier to show the efficacy of our proposed methods. The result is shown in Figure 1.

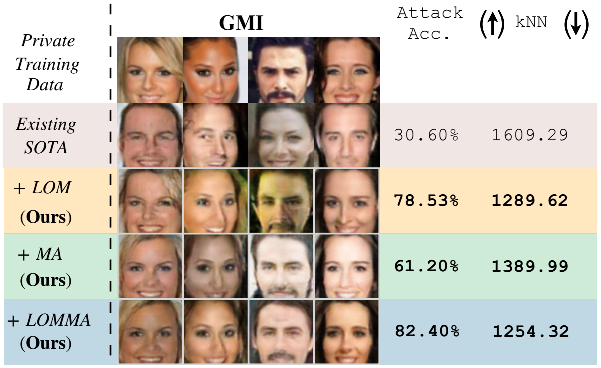

Figure 1. Qualitative / Quantitative (Top1 Attack Acc., KNN Dist) results to demonstrate the efficacy of our proposed method. We use GMI [20], $\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14] and M = IR152 [8]. As one can observe, our proposed method achieves better reconstruction of private data both visually and quantitatively (validated by KNN results) resulting in a significant boost in attack performance.

D.2. Penultimate layer visualization for GMI, KEDMI and VMI

In this section, we show additional penultimate layer visualizations to support our formulation of L_{id}^{logit} as an improved MI Identity Loss. We show visualizations for GMI [20] and VMI [17] in Figures 2 and 5 respectively. Further,

[1] https://github.com/SCccc21/Knowledge-Enriched-DMI

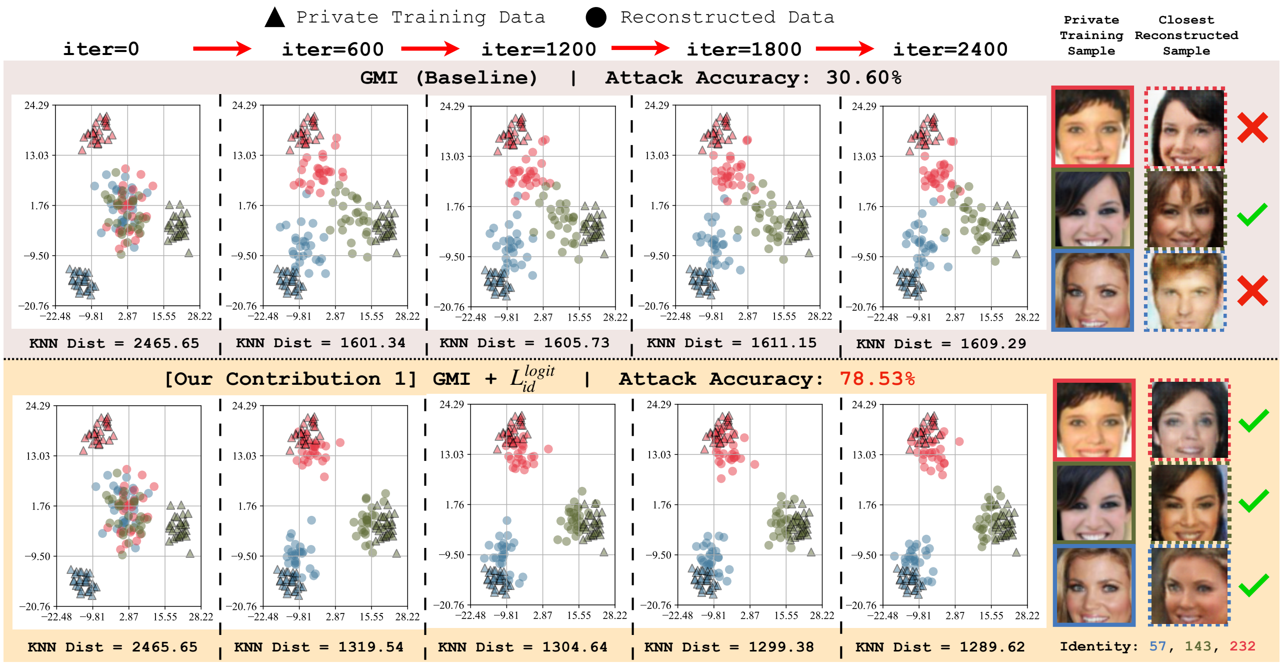

Figure 2. Visualization of the penultimate layer representations ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14], M_t = IR152 [8], Evaluation Model = face.evoLve [2], Inversion iterations = 2400) for private training data and reconstructed data using GMI [20]. Following exact evaluation protocol in [1], we use face.evoLve [2] to extract representations. We show results for 3 randomly chosen identity. We include KNN distance (for different iterations) and final attack accuracy following the protocol in [1]. For each identity, we also include a randomly selected private training data and the closest reconstructed sample at iteration=2400. ① **Identity loss in SOTA MI methods [1, 17, 20] (Eqn. 2, main paper) is sub-optimal for MI (Top).** Using penultimate representations during inversion, we observe 2 instances (*e.g.* target identity **57** and **232**) where GMI [20] (using Eqn. 2, main paper for identity loss) is unable to reconstruct data close to private training data. Hence, private and reconstructed facial images are qualitatively different. ② **Our proposed identity loss, L_{id}^{logit} (Eqn. 3, main paper), can effectively guide reconstruction of data close to private training data (Bottom)**. This can be clearly observed using both penultimate layer representations and KNN distances for all 3 target classes **57**, **143** and **232**. Best viewed in color.

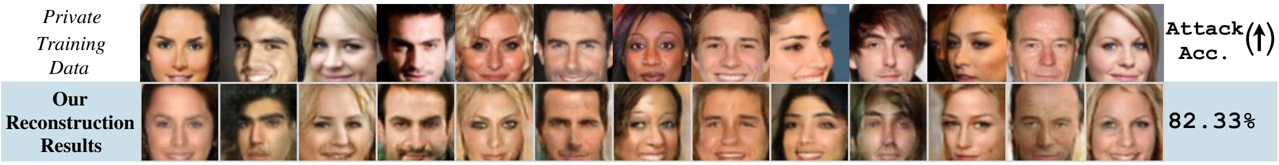

Figure 3. We show private data (top), *our* reconstruction results (bottom) and Attack accuracy ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14], M_t = face.evoLve [2], Evaluation Model = face.evoLve [2], Inversion iterations = 2400) using GMI [20]. We remark that these results are obtained by combining L_{id}^{logit} and L_{id}^{aug} (referred to as **+ LOMMA** throughout the paper).

we show penultimate layer visualization for an additional target classifier, face.evoLve using KEDMI [1] in Figure 8 to validate our findings.

D.3. Our reconstruction results

Given that the goal of MI is to reconstruct private training data, in this section, we show reconstructed samples for 5 additional setups using our proposed method. We show reconstruction results using GMI [20] and VMI [17] in Figures 7 and 9 respectively. Further, we show additional reconstruction results for GMI and KEDMI using a different target classifier (face.evoLve) in Figures 3 and 4 to validate the efficacy of our proposed method. Finally, we show reconstruction results for Cross-dataset MI in Figure 6. We remark that cross-dataset MI is a challenging attack setup due to large distribution shift between private and public data. Following [17], we use FFHQ [13] as the public dataset. To conclude, we remark that the samples reconstructed using our proposed method closely resembles the private training data in many instances, and this is quantitatively validated using MI attack accuracy.

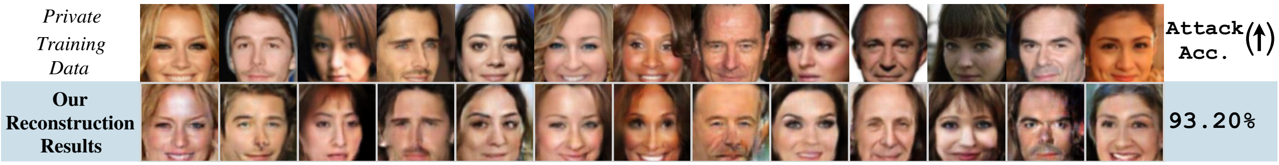

Figure 4. We show private data (top), *our* reconstruction results (bottom) and Attack accuracy ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14], M_t = face.evoLve [2], Evaluation Model = face.evoLve [2], Inversion iterations = 2400) using KEDMI [1]. We remark that these results are obtained by combining L_{id}^{logit} and L_{id}^{aug} (referred to as **+ LOMMA** throughout the paper). We remark that in the standard CelebA benchmark, our method boosts attack accuracy significantly, achieving more than 90% attack accuracy for the first time in contemporary MI literature.

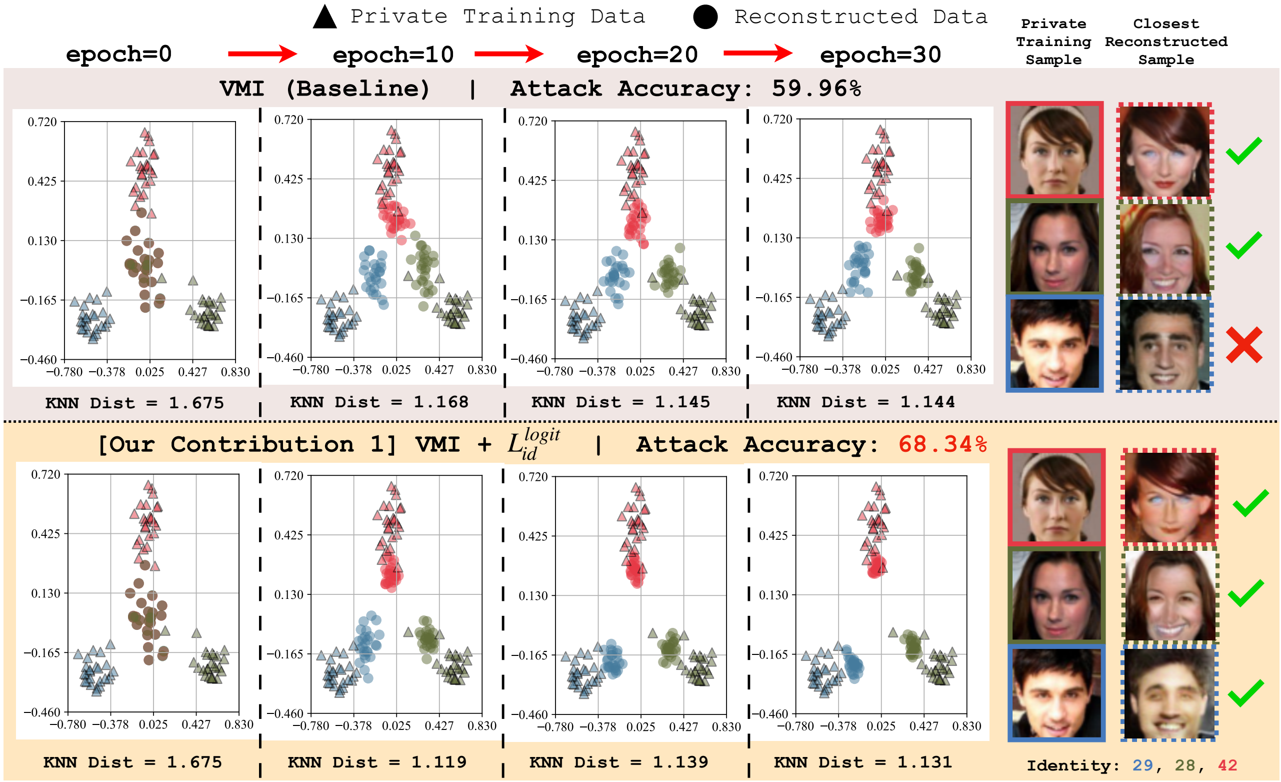

Figure 5. Visualization of the penultimate layer representations ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14], M_t = ResNet34 [17], Evaluation Model = IR-SE50 [17], Inversion epochs = 30) for private training data and reconstructed data using VMI [17]. Following exact evaluation protocol in [17], we use IR-SE50 to extract representations. We show results for 3 randomly chosen identity. We include KNN distance and final attack accuracy. Given that we strictly follow [17], we remark that due to the use of IR-SE50 evaluation classifier to extract penultimate layer representations, the features have different scales resulting in lower KNN distances (compared to KEDMI and GMI results). For each identity, we include a randomly selected private training data and the closest reconstructed sample (epoch = 30). **① Identity loss in SOTA MI methods [1,17,20] (Eqn. 2, main paper) is sub-optimal for MI (Top).** Using penultimate representations during inversion, we observe an instance (*e.g.* target identity 29) where VMI [17] (using Eqn. 2, main paper for identity loss) is unable to reconstruct data close to private training data. Hence, private and reconstructed facial images are qualitatively different. **② Our proposed identity loss, L_{id}^{logit} (Eqn. 3, main paper), can effectively guide reconstruction of data close to private training data (Bottom)**. This can be observed using penultimate layer representations and KNN distances for all 3 target classes 29, 28 and 42. Best viewed in color.

E. Additional Related work

Given a trained model, Model Inversion (MI) aims to extract information about training data. Fredrikson et al. [7] propose one of the first methods for MI. The authors found that attackers can extract genomic and demographic information about patients using the ML model. In [6], Fredrikson et al. extended the problem to the facial recognition setup where the authors can recover the face images. In [19], Yang et al. proposed adversarial model inversion which uses the target classifier as an encoder to produce a

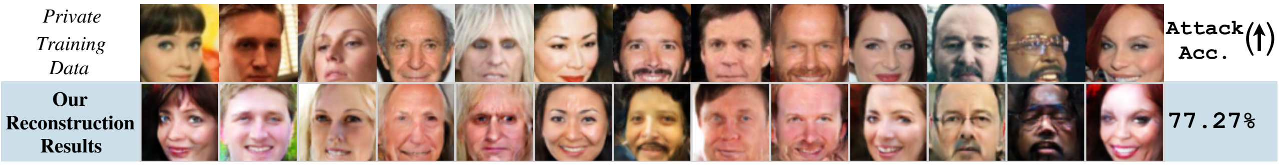

Figure 6. *Cross-dataset MI results*. We show private data (top), *our* reconstruction results (bottom) and Attack accuracy ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = FFHQ [13], M_t = IR152 [8], Evaluation Model = face.evoLve [2], Inversion iterations = 2400) using KEDMI [1]. Cross-dataset MI is a challenging setup due to the large distribution shift between private and public data. We remark that these results are obtained by combining L_{id}^{logit} and L_{id}^{aug} (referred to as **+ LOMMA** throughout the paper).

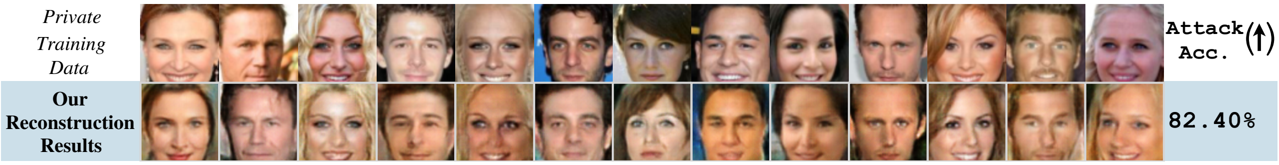

Figure 7. We show private data (top), *our* reconstruction results (bottom) and Attack accuracy ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14], M_t = IR152 [8], Evaluation Model = face.evoLve [2], Inversion iterations = 2400) using GMI [20]. We remark that these results are obtained by combining L_{id}^{logit} and L_{id}^{aug} (referred to as **+ LOMMA** throughout the paper).

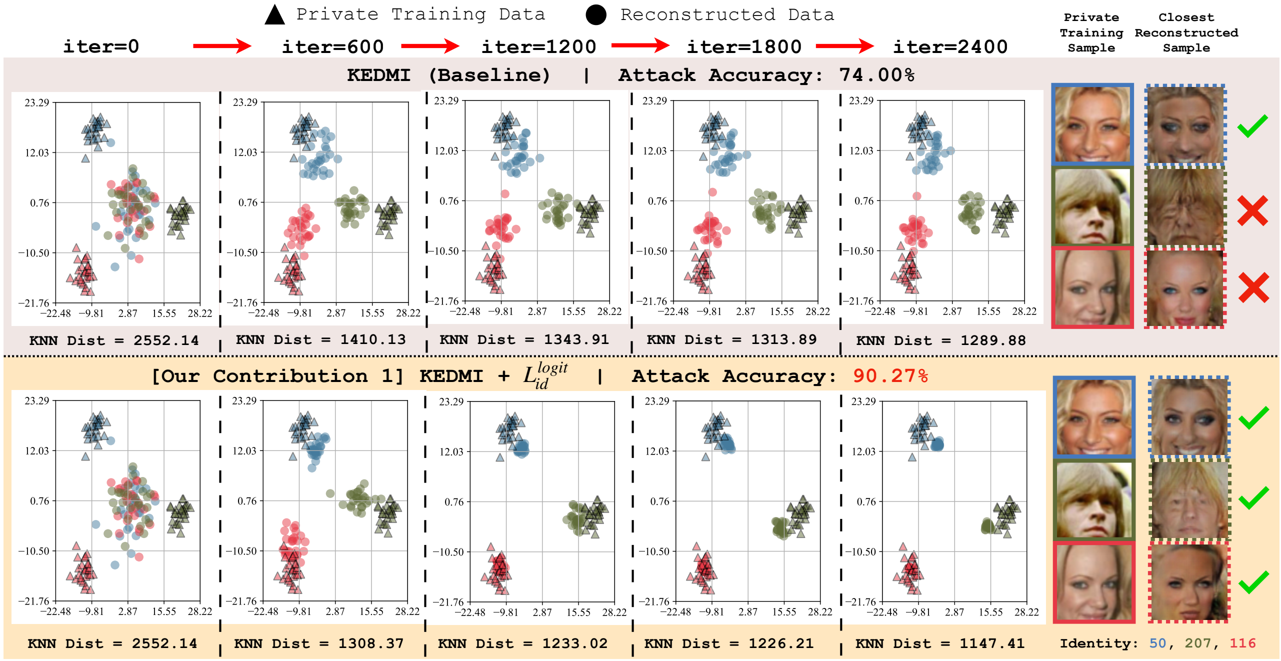

Figure 8. Visualization of the penultimate layer representations ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14], M_t = VGG16 [15], Evaluation Model = face.evoLve [2], Inversion iterations = 2400) for private training data and reconstructed data using KEDMI [1]. Following exact evaluation protocol in [1], we use face.evoLve [2] to extract representations. We show results for 3 randomly chosen identity. We include KNN distance (for different iterations) and final attack accuracy following the protocol in [1]. For each identity, we also include a randomly selected private training data and the closest reconstructed sample at iteration=2400. **① Identity loss in SOTA MI methods [1, 17, 20] (Eqn. 2, main paper) is sub-optimal for MI (Top).** Using penultimate representations during inversion, we observe 2 instances (*e.g.* target identity 207 and 116) where KEDMI [1] (using Eqn. 2, main paper for identity loss) is unable to reconstruct data close to private training data. Hence, private and reconstructed facial images are qualitatively different. **② Our proposed identity loss, L_{id}^{logit} (Eqn. 3, main paper), can effectively guide reconstruction of data close to private training data (Bottom)**. This can be clearly observed using both penultimate layer representations and KNN distances for all 3 target classes 50, 207 and 116. Best viewed in color.

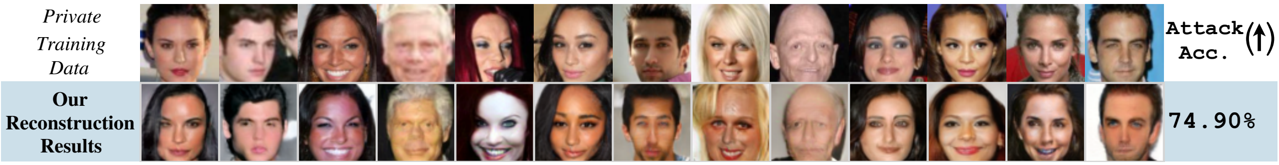

Figure 9. We show private data (top), *our* reconstruction results (bottom) and Attack accuracy ($\mathcal{D}_{priv}$ = CelebA [14], $\mathcal{D}_{pub}$ = CelebA [14], M_t = ResNet34 [17], Evaluation Model = IR-SE50 [5], Inversion epochs = 30) using VMI [17]. We remark that these results are obtained by combining L_{id}^{logit} and L_{id}^{aug} (referred to as **+ LOMMA** throughout the paper).

prediction vector. A second network takes the prediction vector as the input to reconstruct the data.

Instead of performing MI attacks directly on high-dimensional space (e.g. image space), recent works have proposed to reduce the search space to latent space by training a deep generator [1,17,18,20]. In particular, a generator is trained on an auxiliary dataset that has a similar structure to the target image space. In [20], the authors proposed GMI which uses a pretrained GAN to learn the image structure of the auxiliary dataset and finds the inversion images through the latent vector of the generator. Chen et al. [1] extend GMI by training discriminator to distinguish the real and fake samples and to be able to predict the label as the target model. Furthermore, the authors proposed modeling the latent distribution to reduce the inversion time and improve the quality of reconstructed samples. VMI [17] provides a probabilistic interpretation for MI and proposes a variational objective to approximate the latent space of target data.

Zhao et al. [21] propose to embed the information of model explanations for model inversion. A model explanation is trained to analyze and constrain the inversion model to learn useful activations. Another MI attack type is called label-only MI attacks which attackers only access the predicted label without a confidence probability [3, 12]. Recently, Kahla et al. [12] propose to estimate the direction to reach the target class's centroid for an MI attack. In this work, we instead focus on a different problem and propose two improvements to the identity loss which is common among all SOTA MI approaches.

References

[1] 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.

[2] 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 *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1924–1932, 2017.

[3] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pages 1964–1974. PMLR, 2021.

[4] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[6] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[7] 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 (USENIX Security 14)*, pages 17–32, 2014.

[8] 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.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[11] 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.

[12] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15045–15053, 2022.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[14] 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.

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

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

[17] 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.

[18] Ziqi Yang, Ee-Chien Chang, and Zhenkai Liang. Adversarial neural network inversion via auxiliary knowledge alignment. *arXiv preprint arXiv:1902.08552*, 2019.

[19] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

[20] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261, 2020.

[21] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–692, 2021.