Model Inversion Robustness: Can Transfer Learning Help? Supplementary

Sy-Tuyen Ho[1] Koh Jun Hao[1]
Keshigeyan Chandrasegaran[2]† Ngoc-Bao Nguyen[1] Ngai-Man Cheung[1]
[1]Singapore University of Technology and Design (SUTD) [2]Stanford University

hosy_tuyen@sutd.edu.sg ngaiman_cheung@sutd.edu.sg

Overview

In this Supp., we provide additional experimental results, additional analysis, limitation of existing MI defenses, detailed experiment setting, detailed reproducibility, and qualitative results. These are not included in the main paper due to the space limitation. The PyTorch code, a demonstration, inverted data, and pre-trained models are available at our project page: https://hosytuyen.github.io/projects/TL-DMI

Contents

A. Additional Results

A.1. Additional result on BREPMI

Defense	Acc $\Uparrow$	AttAcc $\Downarrow$	Δ $\Uparrow$	KNN $\Uparrow$
No Def.	89.00	69.67	-	1337.01
BiDO	80.35	39.73	3.46	1534.48
TL-DMI	83.41	42.00	**4.90**	1517.38

Table 1. Empirical results for BREPMI [10]. Following the exact experimental setups from BREPMI, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, evaluation model = FaceNet, and target classifier T = VGG16, there are a total of 300 attacked classes. Our proposed TL-DMI achieves better MI robustness, which is quantified by **MI robustness is quantified by the Δ, the ratio of drop in attack accuracy to drop in natural accuracy**

A.2. Additional Empirical Validation on GMI

Beside the empirical validation on VGG16 with KEDMI and ResNet-18 with PPA presented in the main manuscript. We also provide additional empirical validation on VGG16 with GMI in Fig. 1. The observation is consistent with the results in the main manuscript.

Defense	Acc $\Uparrow$	AttAcc $\Downarrow$	Δ $\Uparrow$
No Def.	90.55	83.87	-
TL-DMI	85.60	19.25	**13.05**

Table 2. We follow the MI setup from MIRROR, where T = ResNet-34, $\mathcal{D}_{priv}$ = Stanford Cars, $\mathcal{D}_{pub}$ = LSUN Cars, $\mathcal{D}_{pretrain}$ = ImageNet1K.

† Work done while at SUTD.

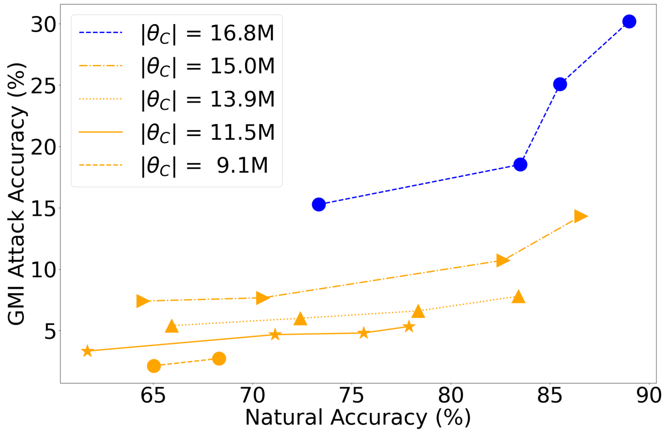


Figure 1. Empirical Validation on VGG16 with GMI. Each line represents one training setup for T with a different $|\theta_C|$ updated on $\mathcal{D}_{priv}$. Note that number of parameters for the entire target model $|\theta_T|$ = 16.8M for this MI setup. To separate the influence of natural accuracy on MI attack accuracy, we perform GMI attacks on different checkpoints for each training setup, varying a wide range of natural accuracy. This is presented by multiple data points on each line. For a given natural accuracy, it can be clearly observed that attack accuracy can be reduced by decreasing $|\theta_C|$, i.e., decreasing parameters updated on $\mathcal{D}_{priv}$.

A.3. Additional result on LOMMA

Due to the remarkably simple implementation of our proposed TL-DMI, we expand the MI robustness evaluation to LOMMA [14], is the SOTA MI attacks. Note that this MI attack has not been included yet in SOTA MI defense BiDO. The results in Tab. 6 shown that TL-DMI is able to defense against SOTA MI attack LOMMA. For a fair comparison, we strictly follow LOMMA for the MI setups.

A.4. Additional result on Stanford Cars dataset

We further show the effectiveness of our proposed TL-DMI on Dogs breeds classification (see Tab. 4) and additional Cars Classification in Tab. 2. The results further show the effectiveness of TL-DMI.

A.5. Additional results on other MI setups

We further provide 3 more setups with PPA in Tab. 4 in this rebuttal. All results consistently support outstanding defense trade-off with TL-DMI

A.6. Comparison with SOTA MI Defense

We provide a comprehensive comparisons between our proposed TL-DMI and BiDO and MID under **6 attacks**: VMI, LOMMA, PPA, KEDMI, GMI, and BREPMI. To avoid the effect of randomness in our comparison, we calculate Δ under 3 attacks of different random seeds. We summarize the comparison in Tab. 5. **All results consistently support that**

Attack	**Defense**	**Acc** ⇑	**AttAcc** ⇓	**Δ** ⇑
PPA	No Def.	94.86	82.83 ± 0.17	-
	MID (0.05)	90.85	52.09 ± 0.45	7.67 ± 0.09
	MID (0.02)	91.54	61.67 ± 0.33	5.77 ± 0.06
	MID (0.01)	92.70	75.84 ± 0.60	2.11 ± 0.15
	TL-DMI	90.10	31.70 ± 0.17	**10.74 ± 0.05**
LOMMA	No Def.	89.00	93.68 ± 1.94	-
	MID (0.002)	78.06	76.51 ± 0.27	1.57 ± 0.20
	MID (0.003)	75.83	78.73 ± 1.88	0.28 ± 0.29
	MID (0.004)	72.87	76.14 ± 0.90	1.09 ± 0.10
	TL-DMI	83.41	72.47 ± 2.85	**3.79 ± 0.19**

Table 3. Varying MID hyperparameters, we conduct three comparisons, reporting mean and standard deviation.

Attack	**T**	**Defense**	**Acc** ⇑	**AttAcc** ⇓	**Δ** ⇑
PPA	ResNet-18	No Def.	94.22	47.41 ± 0.18	-
		TL-DMI	91.12	5.19 ± 0.23	**13.62 ± 0.04**
	ResNet-101	No Def.	96.57	38.99 ± 0.19	-
		TL-DMI	93.01	6.66 ± 0.12	**9.08 ± 0.03**
	MaxViT	No Def.	96.57	31.79 ± 0.20	-
		TL-DMI	93.01	4.11 ± 0.16	**7.78 ± 0.03**

Table 4. Additional MI setups for PPA, where $\mathcal{D}_{priv}$ = FaceScrub, $\mathcal{D}_{pub}$ = Metfaces, $\mathcal{D}_{pretrain}$ = ImageNet1K.

TL-DMI outperforms BiDO and MID

For BiDO reproducibility, we follow the exact hyperparemeters from their work. Note that BiDO is the best defense by far, but it requires extensive grid-search for hyperparameters. For MID reproducibility, we adopt their implementation and hyperparameters. Furthermore, we provide the results for MID with different hyperparameter choices in Tab. 3.

B. Additional Analysis

B.1. The effect of pretrain dataset to MI robustness

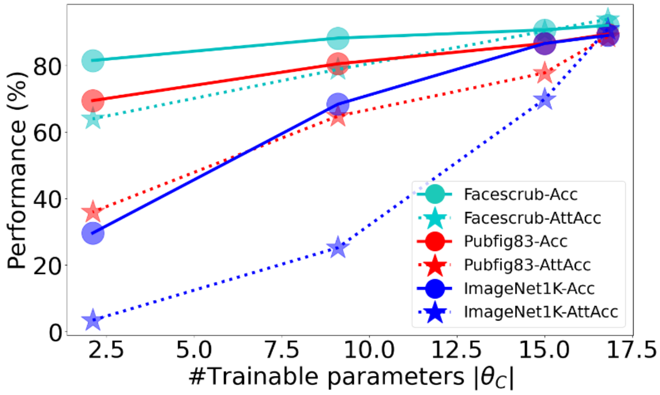
In these above sections, we use a consistent and standard pre-trained dataset to ensure fair comparison with other methods in the literature. Since the pre-trained backbone can be produced with different datasets in practice, we investigate the impact of different pre-trained datasets on MI robustness in this section. Specifically, we implement the same setup as the KEDMI setup for VGG16, but vary three different pre-trained datasets: ImageNet1K, Facescrub, and Pubfig83. The results are shown in Fig. 2.

Updating all parameters $|\theta_C|$ = 16.8M on $\mathcal{D}_{priv}$, yields no significant differences among different $\mathcal{D}_{pretrain}$. This is expected and align with our understanding, where all the parameters in T are exposed to private data during the training of T. With fewer trainable parameters on $\mathcal{D}_{priv}$, we notice clearer differences. Overall, pre-training on a closer domain (Pubfig83 and Facescrub) restores natural accuracy much better than pre-training on a general domain (Ima-

Attack	T	Defense	Acc ⇑	AttAcc ⇓	Δ ⇑
LOMMA	VGG-16	No Def.	89.00	93.68 ± 1.94	-
		MID	78.06	76.51 ± 0.27	1.57 ± 0.20
		BiDO	80.35	66.22 ± 3.74	3.17 ± 0.29
		TL-DMI	83.41	72.47 ± 2.85	**3.79 ± 0.19**
PPA	ResNet-18	No Def.	94.22	90.08 ± 1.40	-
		MID	88.27	48.81 ± 0.28	6.94 ± 0.22
		BiDO	91.33	76.65 ± 0.09	4.65 ± 0.46
		TL-DMI	91.12	21.32 ± 0.90	**22.18 ± 0.74**
	ResNet-101	No Def.	94.86	82.83 ± 0.17	-
		MID	90.85	52.09 ± 0.45	7.67 ± 0.09
		BiDO	90.32	67.43 ± 0.36	3.39 ± 0.09
		TL-DMI	90.10	31.70 ± 0.17	**10.74 ± 0.05**
KEDMI	VGG-16	No Def.	89.00	87.71 ± 2.73	-
		MID	78.06	66.64 ± 0.78	1.93 ± 0.30
		BiDO	80.35	39.77 ± 5.60	5.54 ± 0.33
		TL-DMI	83.41	51.64 ± 1.97	**6.45 ± 0.59**
BREP-MI	VGG-16	No Def.	89.00	70.56 ± 1.84	-
		MID	78.06	16.47 ± 1.07	4.91 ± 0.26
		BiDO	80.35	39.35 ± 0.90	3.61 ± 0.16
		TL-DMI	83.41	41.22 ± 0.69	**5.24 ± 0.41**
VMI	ResNet-34	No Def.	69.27	39.40	-
		MID	52.52	29.05	0.62
		BiDO	61.14	30.25	1.13
		TL-DMI	62.20	21.73 ± 2.08	**2.5 ± 0.30**
GMI	VGG-16	No Def.	89.00	31.25 ± 1.04	-
		MID	78.06	28.78 ± 1.24	0.25 ± 0.16
		BiDO	80.35	6.31 ± 0.54	2.88 ± 0.15
		TL-DMI	83.41	8.47 ± 0.58	**4.08 ± 0.11**

Table 5. We re-run three comparisons, presenting mean and standard deviation. Following VMI setup from BiDO, we encounter code reproducibility issues, and we take the best result reported in BiDO paper.

genet1K).



	#Images	Domain	#Classes
ImageNet1K	1.3M	General Domain	1000
Pubfig83	13K	Facial Domain	83
Facescrub	106K	Facial Domain	530

Figure 2. The effect of different $\mathcal{D}_{pretrain}$, i.e., ImageNet1K, Pubfig83, and Facescrub. We use T = VGG16, $\mathcal{D}_{priv}$ = CelebA. The results suggest that the less similarity between pretrain and private dataset domains can improve defense effectiveness.

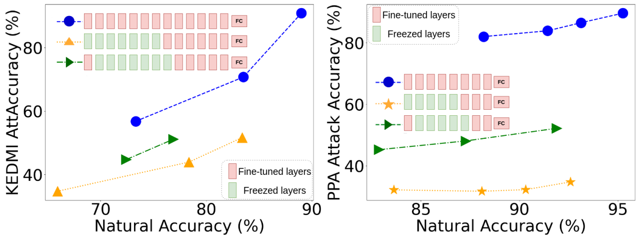


Figure 3. We follow KEDMI-VGG16 and PPA-ResNet-18 setups in Fig. 1-II. Fine-tuning first layers (green line), rather than middle layers (orange line), enhances MI attack accuracy, corroborating our analysis: first layers are important for MI.

For instance, with $|\theta_C|$ = 2.1M, pre-training on Facescrub and Pubfig83 achieve 81.48% and 69.41% accuracy, respectively, compared to 29.59% in the Imagenet1K setup. Nevertheless, pre-training on a closer domain also increases the risk of MI attack. As those frozen parameters during the fine-tuning on $\mathcal{D}_{priv}$ keep the feature representations from $\mathcal{D}_{pretrain}$, thus, the closer the $\mathcal{D}_{pretrain}$, the riskier it is for the model against MI attack. Notably, with $|\theta_C|$ = 15.0M, models pre-training on ImagNet1K and Pubfig83 achieve comparable accuracy. However, using ImageNet1K as $\mathcal{D}_{pretrain}$ renders a more robust model (decreasing MI attack accuracy by 8.06%) than the setup of Pubfig83. In conclusion, when using our TL-DMI to train a MI robust model, it is critical to choose the $\mathcal{D}_{pretrain}$ for a trade-off between restoring model utility and robustness. Specifically, **less similarity between pretrain and private dataset domains can improve defense effectiveness.**

B.2. Layer-wise MI Vulnerability Analysis

we conduct the following experiments which strongly corroborate our analytical results. Specifically, instead of fine-tuning the middle layers, we fine-tune the first layers, see Fig. 3 in this rebuttal. This single change significantly degrades the defense performance and helps MI attacks, which corroborate our analytical results: first layers are important for MI based on our Fisher Information analysis; therefore, fine-tuning the first layers with private dataset helps MI attacks significantly. As another detail to further corroborate our analysis, we remark that first layers have less parameters than middle layers. Yet, MI attacks perform better with fine-tuning private dataset in first layers. This further supports first layers are important for MI. We remark that last layers are critical for classification task, consistent with TL literature. The natural accuracy is much degraded if fine-tuning of last layers is removed.

B.3. Additional Analysis of Layer Importance

FI across MI iterations. MI is a multiple iteration process. The FI for MI in the main manuscript is computed at the last iteration (the iteration that we present the result throughout our submission). Fig. 5 also provides the FI across multiple

Attack Method	$\mathcal{D}_{priv}$	$\mathcal{D}_{pub}$	$\mathcal{D}_{pretrain}$	T	Defense Method	$\lvert\theta_C\rvert/\lvert\theta_T\rvert$	Acc ⇑	Top1-AttAcc ⇓	Top5-AttAcc ⇓	KNN Dist ⇑
LOMMA-K	CelebA	CelebA	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	95.67 ± 0.91	96.68 ± 0.01	1158
					TL-DMI	13.9/16.8	83.41	**75.67 ± 1.83**	**91.68 ± 0.01**	**1304**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	96.40 ± 0.51	99.67 ± 0.15	1038
					TL-DMI	17.8/62.6	86.70	**77.73 ± 1.57**	**94.67 ± 0.66**	**1305**
				FaceNet64	No Def.	35.4/35.4	88.50	89.33 ± 1.19	98.67 ± 0.15	1226
					TL-DMI	34.4/35.4	83.41	**79.60 ± 1.78**	**97.00 ± 0.61**	**1345**
	CelebA	FFHQ	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	58.60 ± 1.67	86.00 ± 1.14	1390
					TL-DMI	13.9/16.8	83.41	**36.00 ± 1.28**	**65.00 ± 1.95**	**1550**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	73.47 ± 1.30	90.00 ± 0.85	1290
					TL-DMI	17.8/62.6	86.70	**45.27 ± 1.98**	**74.33 ± 1.25**	**1474**
				FaceNet64	No Def.	35.4/35.4	88.50	70.27 ± 1.63	90.33 ± 0.72	1391
					TL-DMI	34.4/35.4	83.41	**19.53 ± 1.19**	**41.33 ± 1.34**	**1759**
LOMMA-G	CelebA	CelebA	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	56.00 ± 3.65	79.00 ± 3.84	1454
					TL-DMI	13.9/16.8	83.41	**22.00 ± 4.77**	**45.33 ± 9.08**	**1709**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	64.67 ± 5.54	86.00 ± 5.09	1401
					TL-DMI	17.8/62.6	86.70	**41.87 ± 5.37**	**70.67 ± 5.97**	**1551**
				FaceNet64	No Def.	35.4/35.4	88.50	60.00 ± 5.90	80.00 ± 3.81	1501
					TL-DMI	34.4/35.4	83.41	**43.67 ± 5.60**	**65.00 ± 6.82**	**1616**
	CelebA	FFHQ	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	27.00 ± 6.10	52.33 ± 5.82	1642
					TL-DMI	13.9/16.8	83.41	**8.87 ± 3.12**	**24.00 ± 5.50**	**1829**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	45.20 ± 4.30	70.67 ± 4.58	1503
					TL-DMI	17.8/62.6	86.70	**22.87 ± 5.05**	**43.67 ± 7.46**	**1650**
				FaceNet64	No Def.	35.4/35.4	88.50	30.60 ± 5.21	62.00 ± 5.69	1625
					TL-DMI	34.4/35.4	83.41	**9.33 ± 4.55**	**24.33 ± 4.55**	**1909**

Table 6. Our extended MI robustness evaluation on SOTA MI attack LOMMA [14]. The results of AttAcc and Acc are given in %. We reports the MI defense results against different LOMMA attack setups including LOMMA+KEDMI (LOMMA-K) and LOMMA+GMI (LOMMA-G) with the varying in different public datasets $\mathcal{D}_{pub}$ (CelebA and FFHQ), and pre-trained datasets $\mathcal{D}_{pretrain}$ (Imagenet1K and MS-CelebA-1M).

iterations. We observe that after a few iterations, the FI for earlier layers keeps dominant compared to the later layers.

Different MI losses. In the main manuscript, we use l_2 distance to compute the MI loss. In addition, we provide FI results using l_1 distance and LPIPS [25] to compute the MI loss. The FI results obtained using different MI loss functions are consistent with our main FI observation in the main manuscript.

These additional FI results are consistent with those in our main FI observation in the main manuscript.

B.4. MI Robustness via the False Positive Concept

We provide additional analysis in this Appendix to provide a clear understanding of how our proposed TL-DMI effectively defends against MI attacks, leading to more false positive during MI attacks and decrease in attack accuracy.

As discussed, it has been shown that when a deep neural network-based classifier, denoted as $T = C \circ E$, is pre-trained on a large-scale dataset $\mathcal{D}_{pretrain}$, the features learned in the earlier layers E are transferable to another somewhat related classifier on datasets $\mathcal{D}_{priv}$, enabling the model to maintain its natural accuracy without explicitly updating its parameters on D_{priv} in the earlier layers [23]. This transferability of features benefits our proposed TL-DMI through maintaining the model classification performance and natural accuracy.

In contrast, MI attacks require accurate features to reconstruct the private training dataset $\mathcal{D}_{priv}$. By refraining from updating E on $\mathcal{D}_{priv}$, we limit the leakage of private features into E, thereby improving MI robustness. Specifically, recall MI attacks are usually formulated as:

$$w^* = \arg\min_{w}(-\log P_T(y|G(w)) + \lambda\mathcal{L}_{prior}(w)) \quad (1)$$

Therefore, MI attacks aim to seek w with high likelihood $P_T(y|G(w))$. We make this key observation to un-

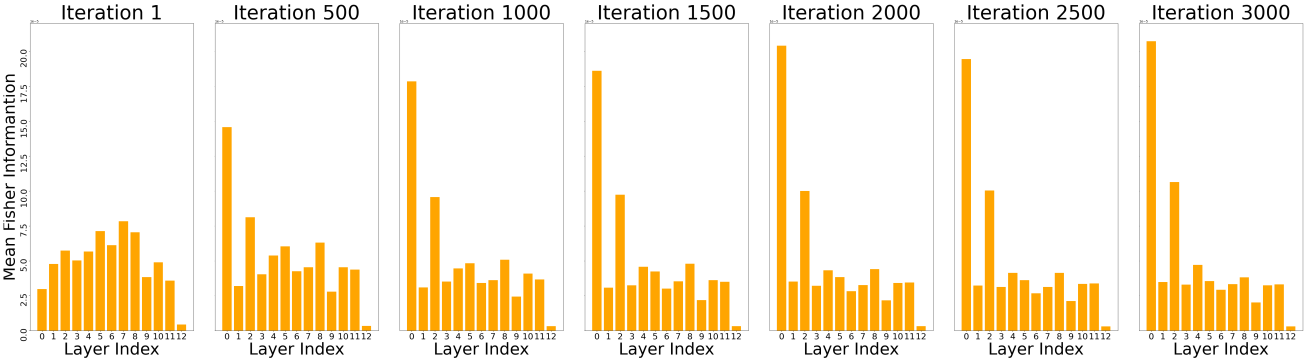


Figure 4. FI distributions across layers during all MI steps. We conduct FI analysis on the main setup in Peng et al [15] where the MI attack is KEDMI [3], T=VGG16, $\mathcal{D}_{priv}$=CelebA and $\mathcal{D}_{pub}$=CelebA. In the main manuscript, we present the FI analysis at the last MI iteration, i.e., iteration 3000. This figures present a more comprehensive FI analysis across multiple iterations. After first few iterations, we consistently observe that the earlier layers are more important to MI task.

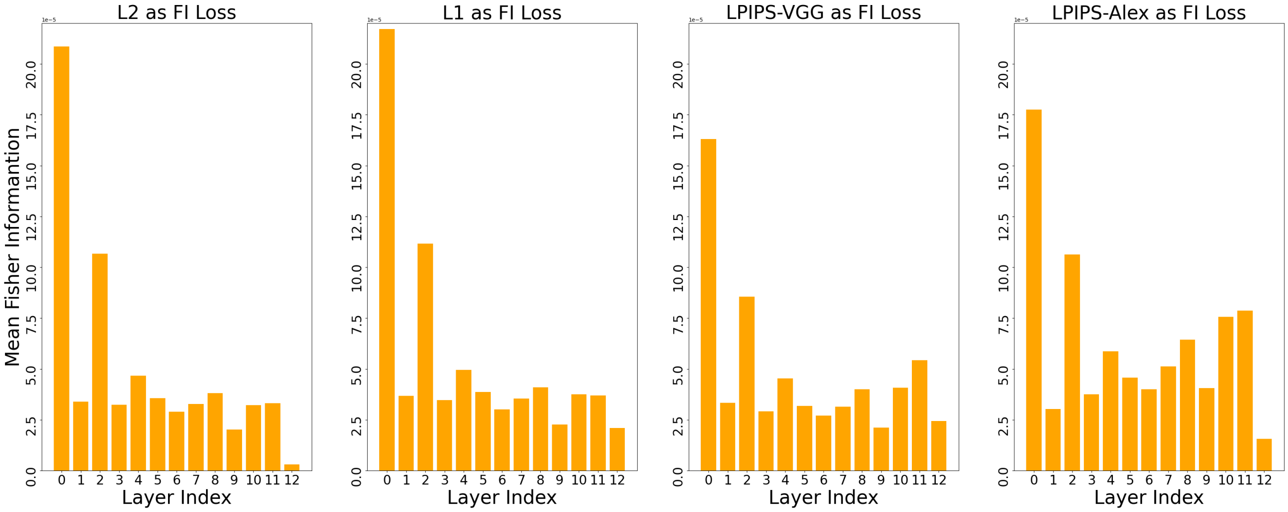
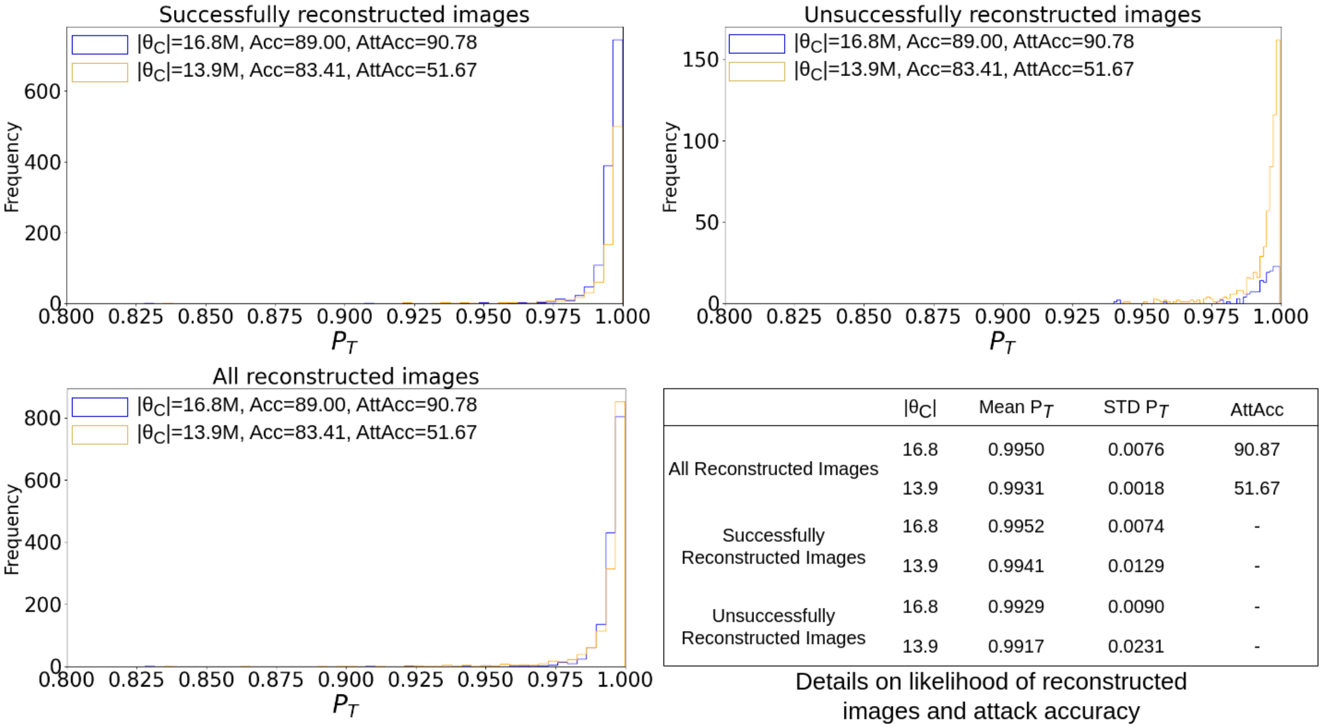


Figure 5. FI distributions across layers via different FI losses. We conduct FI analysis on the main setup in Peng et al [15] where the MI attack is KEDMI [3], T=VGG16, $\mathcal{D}_{priv}$=CelebA and $\mathcal{D}_{pub}$=CelebA. In the main manuscript, we use l_2 distance between reconstructed images and private images as MI loss for the FI analysis. This figure presents the FI analysis through other distances including l_1, LPIPS-VGG [25], LPIPS-ALEX [25]. The results show the consistent observation that the earlier layers of a network are more important to MI attacks compared with later layers.

derstand how our proposed TL-DMI can degrade MI task: *With TL-DMI, while latent variables with high likelihood $P_T(y|G(w))$ can still be identified via Eq. 1, many w^* are false positives, i.e. $G(w^*)$ do not resemble private samples. This results in decrease in attack accuracy.* This can be observed from the likelihood distributions $P_{T_{|\theta_C|=16.8M}}$ and $P_{T_{|\theta_C|=13.9M}}$ for both KEDMI (see Fig. 6) and GMI (see Fig. 7), which are similar and close to 1. These findings indicate that with our proposed TL-DMI, Eq. 1 could still perform well to seek latent variables w to maximize the likelihood $P_T(y|G(w))$. However, although likelihood distributions $P_{T_{|\theta_C|=16.8M}}$ and $P_{T_{|\theta_C|=13.9M}}$ are similar under attacks, the attack accuracy of model with $|\theta_C| = 13.9M$ is significantly lower than that with $|\theta_C| = 16.8M$. This suggests that, due to lack of private data information in E in our proposed TL-DMI model $|\theta_C| = 13.9M$, many w^* do not correspond to images resembling private images.

In the setup where $|\theta_C| = 16.8M$, the optimization process causes the latent variables w to converge towards regions that are closer to the private samples. This outcome is



	$\lvert\theta_C\rvert$	Mean P_T	STD P_T	AttAcc
All Reconstructed Images	16.8	0.9950	0.0076	90.87
	13.9	0.9931	0.0018	51.67
Successfully Reconstructed Images	16.8	0.9952	0.0074	-
	13.9	0.9941	0.0129	-
Unsuccessfully Reconstructed Images	16.8	0.9929	0.0090	-
	13.9	0.9917	0.0231	-

Details on likelihood of reconstructed images and attack accuracy

Figure 6. Visualization of the distribution of P_T for two models: the no defense model (with $|\theta_C| = 16.8M$) and TL-DMI proposed approach (with $|\theta_C| = 13.9M$). The visualization is conducted using KEDMI as the attack method, with $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, $\mathcal{D}_{pretrain}$ = Imagenet1K, and T = VGG16. We observe that both our proposed TL-DMI model and the model without defense exhibit similar distributions of P_T. The values of P_T for both successfully and unsuccessfully reconstructed images are very close to 1 in both cases. However, the attack accuracy shows a significant drop from 90.87 to 51.67 when our proposed TL-DMI is applied.

expected since the model possesses richer low-level features from the private dataset D_{priv} in both E and C. Consequently, we observe more true positives after MI optimization, where the likelihood $P_T(y|G(w))$ is well maximized, and the evaluation model successfully classifies them as label y.

In contrast, in the setup where $|\theta_C| = 13.9M$, the lack of low-level features from $\mathcal{D}_{priv}$ in E hinders the optimization process. As a result, we observe a higher number of false positives after MI optimization. Although these instances successfully maximize the likelihood $P_T(y|G(w))$, the evaluation model is unable to classify them as label y correctly. Therefore, this behavior indicates a higher level of robustness against the MI attack.

C. The limitation of Existing MI Defenses

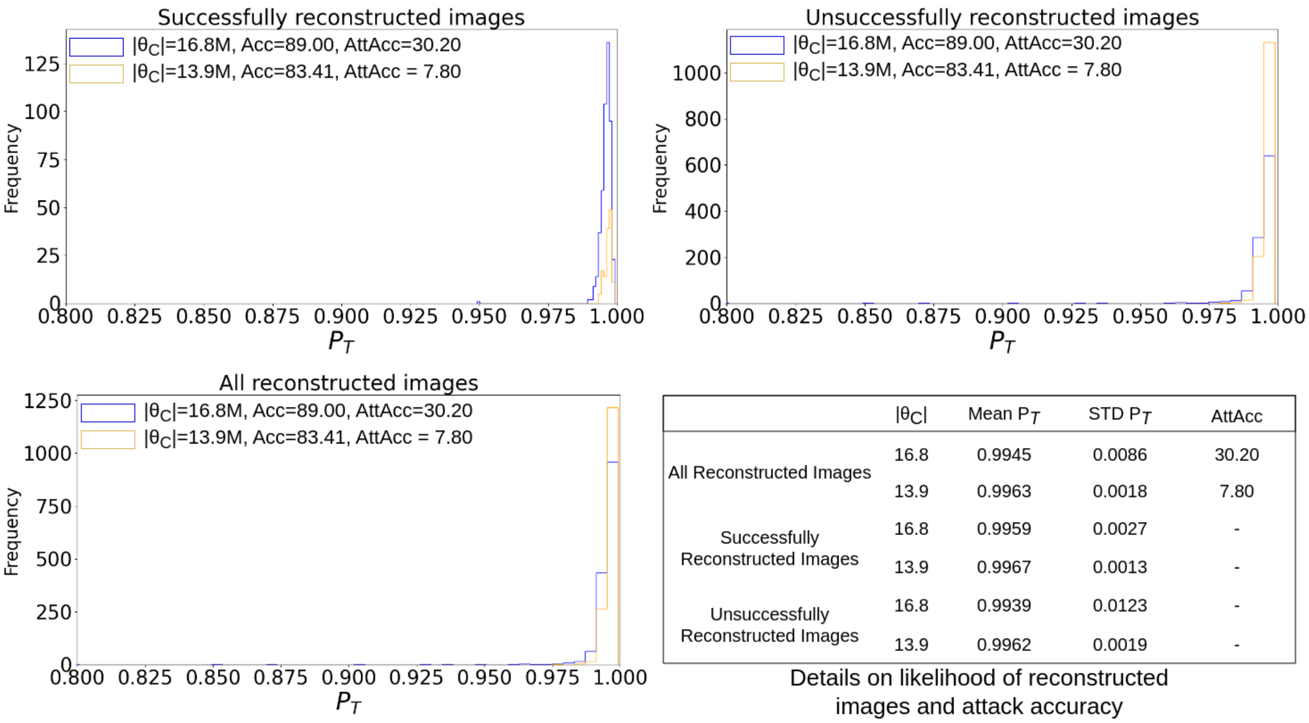
Conflicting objectives between classification and MI defense regularizers: One limitation of the existing MI defenses [15, 22] is the introduction of additional regularizers that conflict with the primary objective of minimizing the classification loss [15]. This conflict often leads to a significant decrease in the overall model utility.

BiDO is sensitive to hyper-parameters. BiDO [15], while attempting to partially recover model utility, suffers from sensitivity to hyper-parameters. Optimizing three objectives simultaneously is a complex task, requiring careful selection of weights to balance the three objective terms. The Tab. 10 results in an explicit accuracy drop when adjusting hyper-parameters λ_x and λ_y even with a small change. The optimized values for λ_x and λ_y in BiDO are obtained through a grid search [15]. For example, in the case of BiDO-HSIC, the authors tested values of $\lambda_x \in [0.01, 0.2]$ and $\frac{\lambda_y}{\lambda_x} \in [5, 50]$. Furthermore, BiDO requires an additional parameter, σ, for applying Gaussian kernels to inputs x and latent representations z in order to utilize COCO [7] and HSIC [6] as dependency measurements.

D. Experiment Setting

D.1. Detailed MI Setup

Attack Dataset. Following existing MI works [3, 14, 21, 26], our work forcuses on the study of CelebA [12]. Furthermore we demonstrate the efficacy of our proposed TL-DMI on other facial datasets with more attack classes (Facescrub [13]) or larger scale (VGGFace2 [2]) and on the an-



	|θC|	Mean P_T	STD P_T	AttAcc
All Reconstructed Images	16.8	0.9945	0.0086	30.20
	13.9	0.9963	0.0018	7.80
Successfully Reconstructed Images	16.8	0.9959	0.0027	-
	13.9	0.9967	0.0013	-
Unsuccessfully Reconstructed Images	16.8	0.9939	0.0123	-
	13.9	0.9962	0.0019	-

Details on likelihood of reconstructed images and attack accuracy

Figure 7. Visualization of the distribution of P_T for two models: the no defense model (with $|\theta_C| = 16.8M$) and TL-DMI proposed approach (with $|\theta_C| = 13.9M$). The visualization is conducted using GMI as the attack method, with $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, $\mathcal{D}_{pretrain}$ = Imagenet1K, and T = VGG16. We observe that both our proposed TL-DMI model and the model without defense exhibit similar distributions of P_T. The values of P_T for both successfully and unsuccessfully reconstructed images are very close to 1 in both cases. However, the attack accuracy shows a significant drop from 90.87% to 51.67% when our proposed TL-DMI is applied.

Architecture	MI Attack	First run		Second run		Third run		Average	
		Natural Acc ⇑	Attack Acc ⇓	Natural Acc ⇑	Attack Acc ⇓	Natural Acc ⇑	Attack Acc ⇓	Natural Acc ⇑	Attack Acc ⇓
VGG16	KEDMI	83.41	51.67 ± 3.93	83.11	49.67 ± 4.86	83.54	53.60 ± 4.06	83.35	51.65 ± 4.28
	GMI		7.80 ± 3.36		8.80 ± 2.28		8.80 ± 3.36		8.47 ± 3.00
Resnet-34	VMI	62.2	23.70 ± 21.38	62.88	19.55 ± 12.90	63.12	21.95 ± 12.36	62.73	21.73 ± 15.91
IR152	KEDMI	86.7	64.60 ± 4.93	86.47	71.6 ± 4.85	86.37	69.33 ± 5.03	86.51	68.51 ± 4.94
	GMI		8.93 ± 3.73		9.47 ± 2.57		9.60 ± 4.16		9.33 ± 3.49
FaceNet64	KEDMI	83.61	73.40 ± 4.10	83.01	76.27 ± 4.09	82.71	76.20 ± 3.96	83.11	75.29 ± 4.05
	GMI		15.73 ± 4.58		15.93 ± 5.20		13.6 ± 3.97		15.09 ± 4.58

Table 7. We present the results for running experiments multiples time to show the reproducibility of our proposed TL-DMI. For KEDMI [3]/GMI [26], we conduct the attacks with $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, $\mathcal{D}_{pretrain}$ = Imagenet1K, and T = VGG16/IR152/FaceNet64. For VMI [21], we conduct the attacks with $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, T = Resnet-34, and there is no $\mathcal{D}_{pretrain}$ for this setup.

imal dataset Stanford Dogs [11]. The details for these datasets used in the experimental setups can be found in Tab. 9.

Attack Data Preparation Protocol. Following previous works [1, 3, 14, 19, 21, 26] approaches, we split the dataset into private $\mathcal{D}_{priv}$ and public $\mathcal{D}_{pub}$ subsets with no class intersection. $\mathcal{D}_{priv}$ is used to train the target classifier T, while $\mathcal{D}_{pub}$ is used to extract general features only.

Target Classifier T. We select VGG16 for T for a fair comparison with SOTA MI defense [15]. As our proposed TL-DMI is architecture-agnostic, we also extend the defense results on more common and recent architectures: i.e., IR152 [9], FaceNet64 [4], Resnet-34, Resnet-18, Resnet-50 [9], ResNeSt-101 [24], and MaxViT [20], which are not explored in previous MI defense setups [15, 22].

Pre-trained Dataset for Target Classifier $\mathcal{D}_{pretrain}$. We use Imagenet-1K [5] for VGG16, Resnet-18/50, ResNeSt-101, and MaxViT, and MS-CelebA-1M [8] for IR152 and FaceNe64, following previous works [3, 26]. For Resnet-34, since it is trained from scratch in the orig-

Architecture	Dataset	Input Resolution	#Epoch	Batch size	Learning rate	Optimizer	Weight Decay	Momentum
VGG16	CelebA	64x64	200	64	0.02	SGD	0.0001	0.9
IR152	CelebA	64x64	100	64	0.01	SGD	0.0001	0.9
FaceNet64	CelebA	64x64	200	8	0.008	SGD	0.0001	0.9
Resnet-34	CelebA	64x64	200	64	0.1	SGD	0.0005	0.9
Resnet-18	CelebA	224x224	100	128	0.001	Adam	-	-
MaxViT	CelebA	224x224	100	64	0.001	Adam	-	-
ResNeSt-101	Stanford Dogs	224x224	100	128	0.001	Adam	-	-
Resnet-50	VGGFace2	224x224	100	1024	0.001	Adam	-	-

Table 8. Training settings for target classifier T. We follow the procedure for training T from previous works [3, 14, 19]

.

	#Classes	#Images	#Attack Classes
CelebA	1,000	27,018	300
Facescrub	530	106,863	530
VGGFace2	8,631	3.31M	100
Stanford Dogs	120	20,580	120

Table 9. MI Private Dataset Setting. We follow previous works [1, 3, 14, 19, 21, 26] for the datasets selection.

λ_x	0.05	0.05	0.05	0.06	1.0	1.0	1.0
λ_y	0.5	0.4	0.6	0.5	0.5	5.0	10.0
Natural Acc	80.35	73.69	76.46	76.13	23.27	57.57	57.04

Table 10. The SOTA MI defense, BiDO is sensitive to hyper-parameters, posing challenges for applying effectively to different architectures of target classifier T or private dataset D_{priv}. BiDO simultaneously optimizes two objectives: $d(x, f)$ (limiting information of input x and feature representations f) and $d(f, y)$ (providing sufficient information about label y to f), in addition to the main objectives $\mathcal{L}$. Therefore, the final objective is $\mathcal{L} + \lambda_x d(x, z) + \lambda_y d(f, y)$, where careful weight selection for λ_x and λ_y is necessary to achieve a balanced training among three objectives. It is clear that inappropriate values of λ_x and λ_y in BiDO cause an unstable training T. *Note that [15] requires an extensive grid search to determine suitable values for λ_x and λ_y*

inal VMI setup [21], we freeze the layers initialized from scratch. In Sec. B, we also study two additional pre-trained datasets, Facescrub [13] and Pubfig83 [16].

MI Attack Method. Our work focuses on white-box attacks, the most effective method in the literature. Following the SOTA MI defense [15], we evaluate our proposed TL-DMI against three well-known attacks: GMI [26], KEDMI [3], and VMI [21]. We further evaluate our proposed TL-DMI against current SOTA MI attacks as well as other SOTA MI attacks LOMMA [14], PPA [19], and MIRROR [1]. The details for MI attack setups can be found below and in the Tab. 11:

- **GMI** [26] uses a pre-trained GAN to understand the image structure of an additional dataset. It then identifies inversion images by analyzing the latent vector of the generator.
- **KEDMI** [3] expands on GMI [26] by training a discriminator to differentiate between real and fake samples and predict the label as the target model. The authors also propose modeling the latent distribution to reduce inversion time and enhance the quality of reconstructed samples.
- **VMI** [21] introduces a probabilistic interpretation of MI and presents a variational objective to approximate the latent space of the target data.
- **LOMMA** [14] introduces two concepts of logit loss for identity loss and model augmentation to improve attack accuracy of previous MI attacks including GMI, KEDMI, and VMI.
- **PPA** [19] proposes a framework for MI attack for high resolution images, which enable the use of a single GAN (i.e., StyleGAN) to attack a wide range of targets, requiring only minor adjustments to the attack.
- **MIRROR** [1] proposes a MI attack framework based on StyleGAN similar to PPA, which aims at reconstructing private images having high fidelity.
- **BREPMI** [10] introduce a new MI attack that can reconstruct private training data using only the predicted labels of the target model. The attacks works by evaluating the predicted labels over a sphere and then estimating the direction to reach the centroid of the target class.

D.2. Evaluation metrics

In the main manuscript, we make use of Natural Accuracy, Attack Accuracy, and K-Nearest-Neighbors Distance (KNN Dist) metrics to evaluate MI robustness. These metrics are described as:

- **Attack accuracy (AttAcc).** To gauge the effectiveness of an attack, we develop an *evaluation classifier* that predicts the identities of the reconstructed images. This metric assesses the similarity between the gener-

	#Iteration	**w clipping**	**Learning rate**	**#Attack per class**
GMI	3000	Yes	0.02	5
KEDMI	3000	Yes	0.02	5
VMI	320	-	0.0001	100
LOMMA	2400	Yes	0.02	5
PPA	50	Yes	0.005	50
MIRROR	500	Yes	0.25	8

Table 11. MI Attack Setups. We follow the MI setups from previous works [1, 3, 14, 15, 19]

Architecture	**Method**	λ_{MID}	λ_x	λ_y	$\lvert\theta_C\rvert$	**Natual Acc** ⇑
VGG16	No. Def	-	-	-	16.8	89.00
	MID	0.01	-	-	-	68.39
	MID	0.003	-	-	-	78.70
	BiDO-COCO	-	10	50	-	74.53
	BiDO-COCO	-	5	50	-	81.55
	BiDO-HSIC	-	0.05	1	-	70.31
	BiDO-HSIC	-	0.05	0.5	-	80.35
	TL-DMI	-	-	-	15.0	86.57
	TL-DMI	-	-	-	13.9	83.41
	TL-DMI	-	-	-	11.5	77.89
	TL-DMI	-	-	-	9.1	69.80
	TL-DMI + BiDO-HSIC	-	0.05	0.4	15.0	84.31
	TL-DMI + BiDO-HSIC	-	0.03	0.4	15.0	82.15
Resnet-34	No. Def	-	-	-	21.5	69.27
	MID	0	-	-	-	52.52
	BiDO-COCO	-	0.05	2.5	-	59.34
	BIDO-HSIC	-	0.1	2	-	61.14
	TL-DMI	-	-	-	21.1	62.20
IR152	No. Def	-	-	-	62.6	93.52
	TL-DMI	-	-	-	17.8	86.70
FaceNet64	No. Def	-	-	-	35.4	88.50
	TL-DMI	-	-	-	34.4	83.61
Resnet-18	No. Def	-	-	-	11.7	95.30
	TL-DMI	-	-	-	8.9	91.17
MaxViT	No. Def	-	-	-	30.9	96.57
	TL-DMI	-	-	-	18.3	93.00
ResNeSt-101	No. Def	-	-	-	48.4	75.07
	TL-DMI	-	-	-	27.9	79.64

Table 12. Hyperparameters setting for training target classifiers. We follow previous work [15] for the hyperparameters selection of MID and BiDO.

ated samples and the target class. If the evaluation classifier attains high accuracy, the attack is considered successful. To ensure an unbiased and informative evaluation, the evaluation classifier should exhibit maximal accuracy.

- **Natural accuracy (Acc).** In addition to assessing the Attack Acc of a released model, it is also necessary to ensure that the model performs satisfactorily in terms of its classification utility. The evaluation of the model's classification utility is typically measured by its natural accuracy, which refers to the accuracy of the model in the classification problem.
- **K-Nearest Neighbors Distance (KNN Dist).** The KNN Dist metric provides information about the proximity between a reconstructed image associated with a particular label or ID, and the images that exist in the private training dataset. This metric is calculated by determining the shortest feature distance between the reconstructed image and the actual images in the private dataset that correspond to the given class or ID. To calculate the KNN Dist, an l_2 distance measure is used between the two images in the feature space, specifically in the penultimate layer of the evaluation model. This distance measure provides insight into the similarity between the reconstructed and the real images in the training dataset for a particular label or ID.
- $\boldsymbol{\delta_{EvalNet}}$ **and** $\boldsymbol{\delta_{FaceNet}}$ These metrics are measured by the squared l_2 distance between the activation in the penultimate layers. $\delta_{EvalNet}$ is computed via Evaluation Model while $\delta_{EvalNet}$ is computed via pre-trained FaceNet [17]. A lower value indicates that the attack results are more visually similar to the training data.
- $\boldsymbol{\ell_2}$ **distance**. ℓ_2 distance measures how similar the inverted images are to the private data by computing the distance between reconstructed features the centroid features of the private data. A lower distance means that the inverted images are more similar to the target class.
- **Frechet inception distance (FID)**. FID is commonly used to evaluate generative model to access the generated images. The FID measures the similarity between two sets of images by computing the distance between their feature vectors. Feature vectors are extracted using an Inception-v3 model that has been trained on the ImageNet dataset. In the context of MI, a lower FID score indicates that the reconstructed images are more similar to the private training images.

Architecture	**Method**	**Total Training Time (Seconds)** ⇓	**Ratio** ⇓	**Natural Acc** ⇑
VGG16	No. Def	2122	1.00	89.00
	BiDO-COCO	3288	1.55	81.55
	BiDO-HSIC	3296	1.55	80.35
	TL-DMI	**1460**	**0.69**	**83.41**
	TL-DMI + BiDO-HSIC	**2032**	**0.96**	**84.14**
IR152	No. Def	6019	1.00	93.52
	TL-DMI	**2808**	**0.47**	**86.70**
FaceNet64	No. Def	16344	1.00	88.50
	TL-DMI	**14448**	**0.88**	**83.61**

Table 13. Computational Resource. We remark that our proposed TL-DMI achieve SOTA MI robustness while reduce the computational cost as we keep the same training protocol and update fewer parameters than No. Def and SOTA MI Defense BiDO.

E. Reproducibility

E.1. The details for training T

Training target classifier $\boldsymbol{T}$. In this work, we employ VGG16 [18], IR152 [9], and FaceNet64 [4] for our investi-

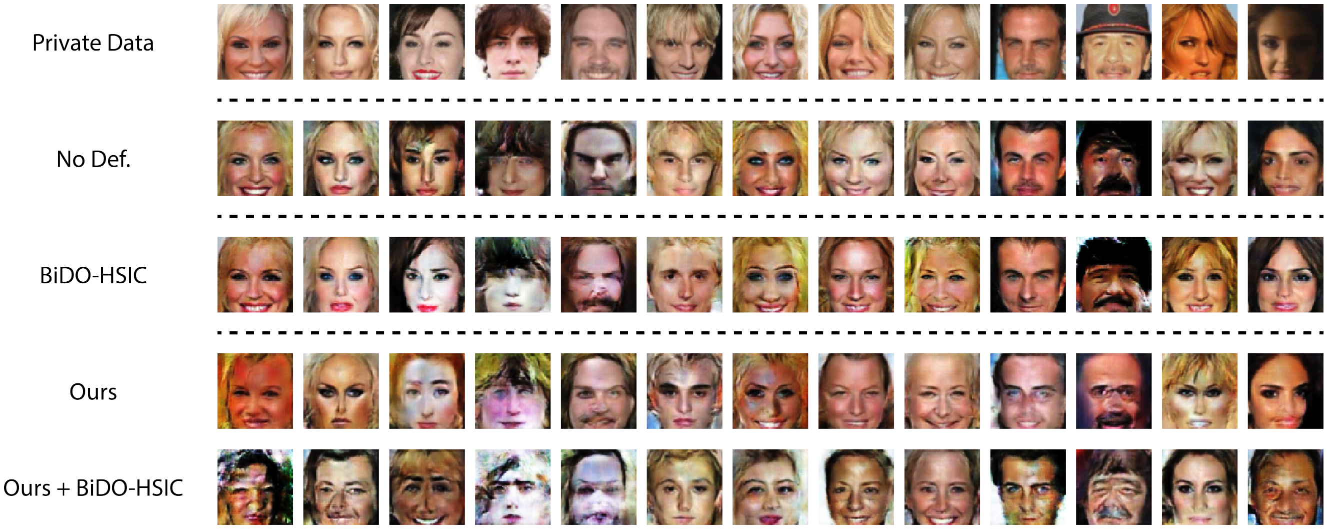


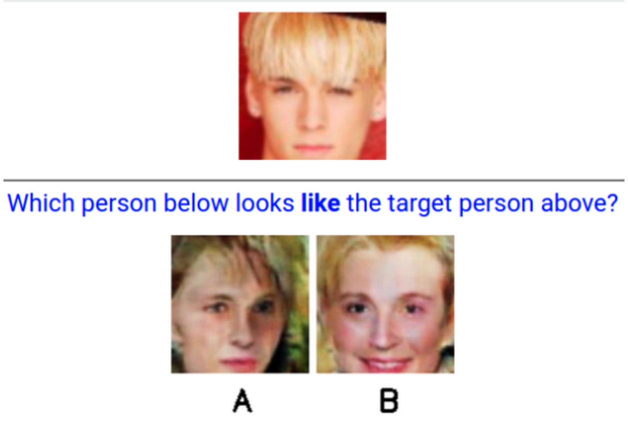
Figure 8. Qualitative results to showcase the effectiveness of our proposed TL-DMI, using KEDMI [3] with $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, $\mathcal{D}_{pretrain}$ = Imagenet1K, and T = VGG16. The visual comparison reveals that our proposed TL-DMI achieves competitive reconstruction of private data, while the hybrid approach combining our method with BiDO-HSIC demonstrates a significant degradation in MI attack and reconstruction quality.

gation. All target classifiers are trained on CelebA dataset. For GMI [26] and KEDMI [3], the target classifiers trained were VGG16, IR152, and FaceNet64, while Resnet-34 was used as the target classifier for VMI [21]. As mentioned in the main manuscript, we employ Imagenet-1K as the pre-trained dataset for VGG16, while MS-CelebA-1M was used as the pre-trained dataset for IR152 and FaceNet64. The details of the training procedure are shown in Tab. 8 below.

Important Hyper-parameters. In our work, we performed an analysis of our proposed TL-DMI against existing SOTA model inversion defense methods: MID [22] and Bilateral Dependency Optimization (BiDO)[15]. MID [22] adds a regularizer $d(x, T(x))$ to the main objective during the target classifier's training to penalize the mutual information between inputs x and outputs $T(x)$. BiDO [15] attempts to minimize $d(x, z)$ to reduce the amount of information about inputs x embedded in feature representations z, while maximizing $d(z, y)$ to provide z with enough information about y to restore the natural accuracy. For simplicity, we use λ_{MID}, λ_x, and λ_y to represent $d(x, T(x))$, $d(x, z)$, and $d(z, y)$ respectively. The settings of these hyper-parameters are detailed in Tab. 12.

E.2. Compute resource

All our experiments are run on NVIDIA RTX A5000 GPUs. Given that our work is focused on model inversion defense, we provide the total training time (seconds) for the target classifier and the ratio of training time between each model inversion defense method against the No. Def. The results in Tab. 13 below show that **our proposed TL-DMI can greatly reduce the amount of time required to train the target classifier.**



Method	NaturalAcc ⇑	AttackAcc ⇓	User Preference ⇓
BiDO	80.35%	45.80%	87.1%
BiDO+TL-DMI	**82.15%**	**17.40%**	**12.9%**

Figure 9. An example for the user study inference (top) and the results for user study (bottom). Compared to BiDO, our proposed TL-DMI provide a better defense with higher natural accuracy but lower user preference.

E.3. Error Bars

For this section, we ran a total of 7 setups (3 times for each setup) across 4 different architectures of the target classifiers, and report their respective natural accuracy and attack accuracy values. For each experiment, we use the same MI attack setup and training settings for target classifiers as reported in the main setups comparing with BiDO and Tab. 8 respectively. We show that the results obtained are reproducible and do not deviate much from the reported values in the main paper. These results can be found in Tab. 7 below.

F. Qualitative results

F.1. Visual Comparison

We evaluate the efficacy of our proposed TL-DMI along with BiDO for preventing privacy leakage on CelebA and also provide visualisation of the samples produced using the KEDMI [3] MI attack method. In Fig. 8 below, each column represents the same identity and the first row represents the ground-truth private data while each subsequent row shows the attack samples reconstructed for each MI defense method.

F.2. User study

We conduct our user study via Amazon MTurk with the interface as shown above. We adapt our user study from MIRROR. In the setup, participants are presented with a real image of the target class, and then asked to pick one of two inverted images that is more closely aligned with the real image. The order is randomized, with each image pair displayed on-screen for a maximum duration of 60 seconds. The assessment encompassed all 300 targeted classes. Each pair of inverted images is assigned to 10 unique individuals, thus our user study involves a total of 3000 pairs of inverted images. We use KEDMI as the MI attack with $\mathcal{D}_{priv} = CelebA$, $\mathcal{D}_{pub} = CelebA$, $T = FaceNet$. *Consistent with the AttackAcc, the user study shows that our proposed TL-DMI provides better defense against the reconstruction of private data characteristics compared to BIDO.*

References

[1] 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. 7, 8, 9

[2] 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. 6

[3] 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. 5, 6, 7, 8, 9, 10, 11

[4] 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. 7, 9

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[6] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16*, pages 63–77. Springer, 2005. 6

[7] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, Bernhard Schölkopf, et al. Kernel methods for measuring independence. 2005. 6

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 7

[9] 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. 7, 9

[10] 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. 1, 8

[11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. 7

[12] 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. 6

[13] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014. 6, 8

[14] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 6, 7, 8, 9

[15] Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *KDD*, 2022. 5, 6, 7, 8, 9, 10

[16] 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. 8

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 9

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 9

[19] 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. 7, 8, 9

[20] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 7

[21] 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. 6, 7, 8, 10

[22] 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. 6, 7, 10

[23] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 4

[24] 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/CVF conference on computer vision and pattern recognition*, pages 2736–2746, 2022. 7

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 5

[26] 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. 6, 7, 8, 10