Model Inversion Robustness: Can Transfer Learning Help?

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

Abstract

Model Inversion (MI) attacks aim to reconstruct private training data by abusing access to machine learning models. Contemporary MI attacks have achieved impressive attack performance, posing serious threats to privacy. Meanwhile, all existing MI defense methods rely on regularization that is in direct conflict with the training objective, resulting in noticeable degradation in model utility. ***In this work,*** *we take a different perspective, and propose a novel and simple Transfer Learning-based Defense against Model Inversion (TL-DMI) to render MI-robust models. Particularly, by leveraging TL, we limit the number of layers encoding sensitive information from private training dataset, thereby degrading the performance of MI attack. We conduct an analysis using Fisher Information to justify our method. Our defense is remarkably simple to implement. Without bells and whistles, we show in extensive experiments that TL-DMI achieves state-of-the-art (SOTA) MI robustness.* ***Our code, pre-trained models, demo and inverted data are available at: https://hosytuyen.github.io/projects/TL-DMI***

1. Introduction

Model Inversion (MI) attack is a type of privacy threat that aim to reconstruct private training data by exploiting access to machine learning models. State-of-the-art (SOTA) MI attacks [6, 32, 33, 43, 51] have demonstrated increased effectiveness, achieving attack performance of over 90% in face recognition benchmarks. The implications of this vulnerability are particularly concerning in security-critical applications [1, 5, 11, 12, 14, 17, 25, 30, 37, 45].

The aim of our work is to propose new perspective to defend against MI attacks and to improve MI robustness. In particular, *MI robustness* pertains to the tradeoff between MI attack accuracy and model utility. MI robustness involves two critical considerations: Firstly, a MI robust model should demonstrate a significant reduction in MI attack accuracy, making it difficult for adversaries to reconstruct private training samples. Secondly, while defending against MI attacks, the natural accuracy of a MI robust model should remain competitive. A model with improved MI robustness ensures that it is resilient to MI while maintaining its utility.

Research gap. Despite the growing threat arising from SOTA MI, there are limited studies on defending against MI attacks and improving MI robustness. Conventionally, differential privacy (DP) is used for ensuring the privacy of individuals in datasets. However, DP has been shown to be ineffective against MI [13, 44, 51]. Meanwhile, a few MI defense methods have been proposed. Particularly, all existing SOTA MI defense methods are based on the idea of *dependency minimization regularization* [35, 44]: they introduce additional regularization into the training objective, with the goal of minimizing the dependency between input and output/latent representation. The underlying idea of these works is to reduce correlation between input and output/latent, which MI attacks exploit during the inversion. However, reducing correlation between input and output/latent directly undermines accuracy of the model, resulting in considerable degradation in model utility [44]. To partially restore the model utility, BiDO [35] proposes to further introduce another regularization to compensate for the reduced correlation between input and latent. However, with two additional regularization along with the original training objective, BiDO requires significant effort in hyperparameter tuning based on intensive grid search [35], and is sensitive to small changes in hyperparameters (see our analysis in Supp.)

In this paper, our main hypothesis is that *a model with fewer parameters encoding sensitive information from private training dataset ($\mathcal{D}_{priv}$) could achieve better MI robustness*. Based on that, we propose a novel Transfer Learning-based Defense against Model Inversion (TL-DMI) (Fig. 1). Leveraging on standard two-stages TL framework [34, 48], with pre-training on public dataset as

† Work done while at SUTD.

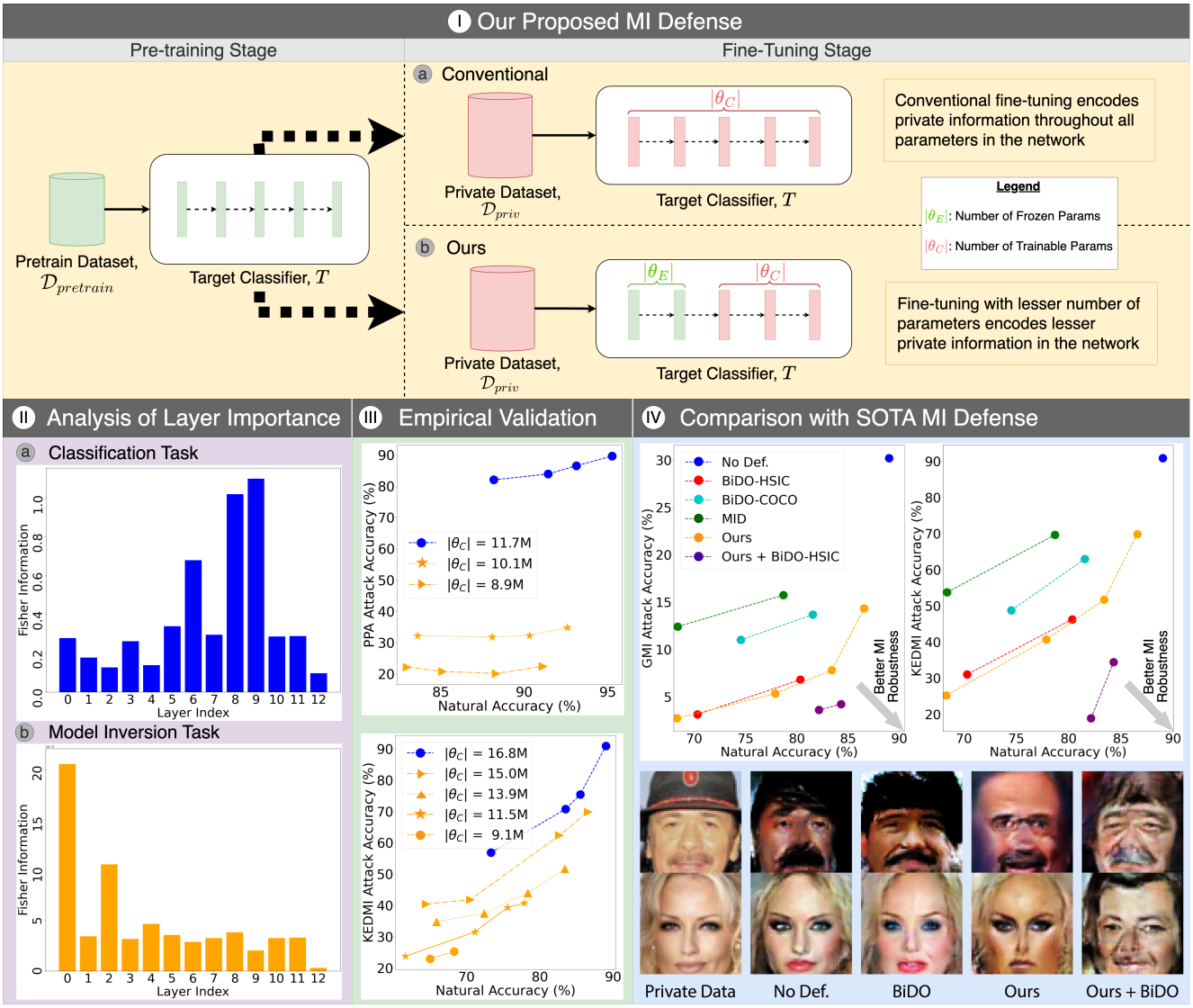

Figure 1. **(I) Our proposed Transfer Learning-based Defense against Model Inversion (TL-DMI) (Sec. 3).** Based on standard TL framework with pre-training (on public dataset) followed by fine-tuning (on private dataset), we propose a simple and highly-effective method to defend against MI attacks. Our idea is to limit fine-tuning with private dataset to a specific number of layers, thereby limiting the encoding of private information to these layers only (pink). Specifically, we propose to perform fine-tuning only on the last several layers. **(II) Analysis of layer importance for classification task and MI task (Sec. 4.2).** For the first time, we analyze importance of target model layers for MI. For a model trained with conventional training, we apply FI and find that the first few layers of the model are important for MI. Meanwhile, FI analysis suggests that last several layers are important for a specific classification task, consistent with TL literature [48]. This supports our hypothesis that preventing the fine-tuning of the first few layers on private dataset could degrade MI significantly, while such impact for classification could be small. Overall, this leads to improved MI robustness. **(III) Empirical validation (Sec. 4.3).** The sub-figures clearly show that at the same natural accuracy, lower MI attack accuracy can be achieved by reducing the number of parameters fine-tuned with private dataset. **(IV) Comparison with SOTA MI Defense (Sec. 4.4).** Without bells and whistles, our method achieves SOTA in MI robustness. Visual quality of MI-reconstructed images from our model is inferior. User study confirms this finding. Extensive experiments can be found in Sec. 4.5. **Best viewed in color with zooming in.**

the first stage and fine-tuning on private dataset as the second stage, we propose to limit private dataset fine-tuning only on a specific number of layers. Specifically, in the second stage, we perform private dataset fine-tuning only on the last several layers of the model. The first few layers are frozen during the second stage, preventing private information encoded in these layers. We hypothesize that by reducing the number of parameters fine-tuned with private dataset, we could reduce the amount of private information encoded in the model, making it more difficult for adver-

saries to reconstruct private training data.

To justify our design, we conduct, for the first time, an analysis of model layer importance for the MI task. We propose to apply Fisher Information (FI) to quantify importance of individual layers for MI [22, 28]. Our analysis suggests that first few layers are important for MI. Therefore, by preventing private information encoded in the first few layers as in our proposed method, we could degrade MI significantly. Meanwhile, during pre-training, the first few layers learn low level information (edges, colour blobs). It is known that low level information is generalizable across datasets [48]. Therefore, our proposed TL-DMI has only small degrade in model utility. Overall, TL-DMI could achieve SOTA MI robustness. We remark that TL-DMI is very easy to implement. In our experiments, we apply TL-DMI to a range of models (CNN, vision transformers), see Sec. 4.5. On the contrary, BiDO has been applied to only VGG16 [40] and ResNet-34 [15]. Our contributions are:

- We propose a simple and highly effective Transfer Learning-based Defense against Model Inversion (TL-DMI). Our idea is a novel and major departure from existing MI defense based on dependency minimization regularization. Furthermore, while majority of TL work focuses on improving model accuracy [18, 34], our work focuses on degrading MI attack accuracy via TL.
- We conduct the first study to analyze layer importance for MI task via Fisher Information. Our analysis results suggest that the first few layers are important for MI, justifying our design to prevent private information encoded in the first few layers.
- We conduct empirical analysis to validate that lower MI attack accuracy can be achieved by reducing the number of parameters fine-tuned with private dataset. Our analysis carefully removes the influence of natural accuracy on MI attack accuracy.
- We conduct comprehensive experiments to show that our proposed TL-DMI achieves SOTA MI robustness. As TL-DMI is remarkably easy to implement, we extend our experiments for a wide range of model architectures such as vision transformer [42], which MI robustness has not been studied before.

2. Background

The target model T is trained on a private training dataset $\mathcal{D}_{priv} = \{(x_i, y_i)_{i=1}^N\}$, where $x_i \in \mathbb{R}^{d_X}$ is the facial image and $y_i \in \{0,1\}^K$ is the identity. The target classifier T is a K-way classifier T: $\mathbb{R}^{d_X} \rightarrow \mathbb{R}^K$, with the parameters $\theta_T \in \mathbb{R}^{d_\theta}$.

Model Inversion (MI) Attack. In MI attacks, an adversary exploits a target model T trained on a private dataset $\mathcal{D}_{priv}$. However, $\mathcal{D}_{priv}$ should not be disclosed. The main goal of MI attacks is to extract information about the private samples in $\mathcal{D}_{priv}$. The existing literature formulates MI attacks as a process of reconstructing an input $\hat{x}$ that T is likely to classify into the preferred class (label) y. This study primarily focuses on whitebox MI attacks, which are the most dangerous, and can achieve impressive attack accuracy since the adversary has complete access to the target model. For high-dimensional data like facial images, the reconstruction problem is challenging. To mitigate this issue, SOTA MI techniques suggest reducing the exploration area to the meaningful and pertinent images manifold using a GAN. Under white-box MI, the adversary can access $T(\hat{x})$, the K-dim vector of soft output, and public dataset $\mathcal{D}_{pub}$ used to train GAN. The Eq. 1 represents the step of existing SOTA white-box MI attacks [3, 6, 32, 41, 51]. The details for SOTA MI attacks can be found in the Supp.

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

where $-\log P_T(y|G(w))$ denotes identity loss in MI attack, which guides the reconstructed $\hat{x} = G(w)$ that is most likely to be classified as class y by T. G refers to generator to generate reconstructed data $\hat{x}$ from latent vector w. The $\mathcal{L}_{prior}$ is the prior loss, which makes use of public information to learn a distributional prior through a GAN. This prior is used to guide the inversion process to reconstruct meaningful images. The hyper-parameter λ is to balance prior loss and identity loss.

Model Inversion (MI) Defense. In contrast, the MI defense aims at minimizing the disclosure of training samples during the MI optimization process. First MI-specific defense strategy is MID [44], which adds a regularization $d(\hat{x}, T(\hat{x}))$ to the main objective during the target classifier's training to penalize the mutual information between inputs $\hat{x}$ and outputs $T(\hat{x})$. Another approach is Bilateral Dependency Optimization (BiDO) [35], which minimizes $d(\hat{x}, f)$ to reduce the amount of information about inputs $\hat{x}$ embedded in feature representations f, while maximizing $d(f, y)$ to provide f with enough information about y to restore the natural accuracy. **However, both MID and BiDO suffer from the drawback that their regularization, i.e., $d(\hat{x}, T(\hat{x}))$ for MID and $d(\hat{x}, f)$ for BiDO, conflict with the main training objective, resulting in an explicit trade-off between MI robustness and model utility**. BiDO improves this trade-off with $d(f, y)$ but is hyperparameter-sensitive due to the optimization of three objectives, making it difficult to apply.

Model inversion (MI) vs. Membership inference. Beside MI, membership inference [16, 23, 36, 38, 39] is another privacy attack on machine learning models. However, **the focus of our work, i.e., vision MI attacks, is fundamentally different from membership inference attacks.** In a membership inference attack, the attacker's objective is to determine whether a specific data point was part of the training dataset used to train the target model. Mem-

	No Defense	Existing MI defenses	Our proposed TL-DMI
Stage 1	Train T with standard objective on $\mathcal{D}_{pretrain}$		
Stage 2	Fine-tune the whole T with standard objective on $\mathcal{D}_{priv}$	Fine-tune the whole T with standard objective and additional *dependency minimization regularization* on $\mathcal{D}_{priv}$	Fine-tune only C with standard objective on $\mathcal{D}_{priv}$

Table 1. Training procedure for "no defense", existing MI defense methods [35, 44] and our proposed TL-DMI. Stage 1 (pre-training) is commonly used in existing methods to reduce the requirement for labeled datasets. TL-DMI takes advantage of such setup to defend MI.

bership inference attacks are typically formulated as a prediction problem, where an attacker model is trained to output *the probability* of a given data point being a member of the training dataset. In contrast, vision model inversion attacks are usually formulated as an image reconstruction problem. The attacker aims to output *the reconstruction* of high-dimensional training images. While membership inference attacks are limited to determining membership status (in or out of the training dataset) and may not provide fine-grained information about the training data, model inversion attacks attempt to recover the training data itself, which can be more invasive [51].

3. Transfer Learning-based Defense against Model Inversion (TL-DMI)

Transfer Learning (TL). TL [34, 47] is an effective approach to leverage knowledge learned from a general task to enhance performance in a different task. By performing pre-training on a large general dataset and then fine-tuning on a target dataset, TL mitigates the demand for large labeled datasets, while simultaneously improving generalization and overall performance. In machine learning, TL works mostly focus on improving the model performance by adapting the knowledge to new tasks and domains [18, 52].

Our proposed defense TL-DMI. In contrast, our work is the first to apply TL to defend against MI attacks aiming at degrading MI attack accuracy. Therefore, our study is fundamentally different from existing TL works which aim to improve model utility [20, 24, 26, 34, 46]. Our idea is to apply TL to reduce the leak of private information by limiting the number of parameters updated on private training data. Specifically, as illustrate in Fig. 1, we propose to train the target model T as $T = C \circ E$ in two stages: pre-training and then fine-tuning. Particularly, in the fine-tuning stage, E comprises parameters that are frozen, i.e., not updated by the private dataset $\mathcal{D}_{priv}$, while C comprises parameters that are updated by $\mathcal{D}_{priv}$.

- **Stage 1: Pre-training with $\mathcal{D}_{pretrain}$.** We first pre-train T using a dataset $\mathcal{D}_{pretrain}$. $\mathcal{D}_{pretrain}$ can be a general domain dataset, e.g., Imagenet1K, or it can be similar domain as the private dataset $\mathcal{D}_{priv}$. Importantly, $\mathcal{D}_{pretrain}$ has no class/identity intersection with $\mathcal{D}_{priv}$. Both C and E are updated based on $\mathcal{D}_{pretrain}$ in this stage.
- **Stage 2: Fine-tuning with $\mathcal{D}_{priv}$.** To adapt the pre-trained model from Stage 1 for $\mathcal{D}_{priv}$, we freeze E, i.e. parameters of E are unchanged. We only update C with $\mathcal{D}_{priv}$.

Tab. 1 provides a comparison between our defense TL-DMI and existing MI defenses. *We remark that pre-training has already been commonly adopted in previous works of MI attack. Therefore, in many cases, our method does not incur additional overhead [3, 6, 32, 35, 41].* As an example, we consider the main setup of BiDO [35] where VGG16 [40] is used as the target classifier T. Following the previous works on MI attack, T including E and C are first pre-trained on $\mathcal{D}_{pretrain}$ = Imagenet1K [10]. Then, for TL-DMI, we fine-tune C with $\mathcal{D}_{priv}$ = CelebA [29] while E is frozen. In contrast, for other MI defense, both E and C are updated with $\mathcal{D}_{priv}$. We explore the design of T with different number of layers updated by $\mathcal{D}_{priv}$, leading to different number of parameters in C ($|\theta_C|$) updated by $\mathcal{D}_{priv}$. Using different $|\theta_C|$, we limit the amount of private information encoded in the parameters of T. We show that our approach TL-DMI improves MI robustness.

Regarding hyperparameter in our proposed TL-DMI, we determine $|\theta_C|$ by simply deciding at the *layer-level* of a deep neural network. Note that during training we use the same objective of classification task, i.e. no change in training objective is needed. Therefore, TL-DMI is much simpler and faster than SOTA MI defense BiDO [35] (see Supp.). **In Sec. 4.2, we present our Fisher Information-based analysis to justify TL-DMI.**

4. Exploring MI Robustness via Transfer Learning

We introduce the experiment setup in Sec. 4.1. In Sec. 4.2, we provide the first analysis on layer importance for MI task via Fisher Information suggesting that earlier layers are important for MI. Then, Sec. 4.3 empirically validate that MI robustness is obtained by reducing the number of parameters fine-tuned with private dataset. With the established understandings, we then compare our proposed method with current SOTA MI defenses [35, 44] in Sec. 4.4. Additionally, since our method offer higher practicality com-

pared with the SOTA MI defenses, we extensively access our approach on 20 MI attack setups in Sec. 4.5 and Supp., spanning 9 architectures, 4 private datasets $\mathcal{D}_{priv}$, 3 public datasets $\mathcal{D}_{pub}$, and 7 MI attacks.

While the above sections assume a consistent pre-trained dataset $\mathcal{D}_{pretrain}$ for the target classifier to ensure fair comparison with existing works, we also delve into novel analysis on the effect of various $\mathcal{D}_{pretrain}$ on MI robustness. We observe that *less similarity between pretrain and private dataset domains can improve defense effectiveness*. The details for this analysis can be found in Supp.

4.1. Experimental Setup

To ensure a fair comparison, our study strictly follows setups in SOTA MI defense method BiDO [35] in datasets, attack methods, and network architectures. Furthermore, we also examine our defense approach with additional new datasets, recent MI attack models, and new network architectures. Note that these have not been included in BiDO. All the MI setups in our study are summarized in Tab. 2. The details for the setup can be found in Supp.

MI Defense Baseline. In order to showcase the efficacy of our proposed TL-DMI, we compare TL-DMI with several existing SOTA model inversion defense methods, which are BiDO [35] and MID [44].

Evaluation Metrics. Following the previous MI defense/attack works, we adopt natural accuracy (Acc), Attack Accuracy (AttAcc), K-Nearest Neighbors Distance (KNN Dist), and ℓ_2 distance metrics to evaluate MI robustness. Moreover, we also provide qualitative results and user study in the Supp.

4.2. Analysis of Layer Importance for Classification Task and MI Task

In this section, we provide an analysis to justify our proposed TL-DMI to render MI robustness. We aim to understand importance of individual layers for MI reconstruction task, justifying our design in TL-DMI to prevent encoding of private data information in the first few layers as an effective method to degrade MI. We study layer importance between classification and MI tasks. To quantify the importance, we compute the Fisher Information (FI) for the two tasks for individual layers.

Fisher Information (FI) based analysis. Fisher Information F has been applied to measure the importance of model parameters for discriminative task [2, 22] and generative task [28]. For example, in [22], FI has been applied to determine the importance of model parameters to overcome the catastrophic forgetting in continual learning. Our study extends FI-based analysis for model inversion, which has not been studied before. Specifically, given a model T parameterized by θ_T and input X, FI can be computed as

Attack Method	$\mathcal{D}_{pub}$	$\mathcal{D}_{priv}$	T
VMI [43]	CelebA [29]	CelebA [29]	ResNet-34 [15]
KEDMI[6]/ GMI [51]			VGG16 [40]
LOMMA [32] / BREPMI [19]			
KEDMI [6] / GMI [51]	CelebA [29] / FFHQ [21]	CelebA [29]	FaceNet64[7]/ IR152 [15]
	FFHQ [21]		VGG16 [40]
PPA [41]	FFHQ [21]	FaceScrub [31]	ResNet-18 [15] / ResNet-101 [15] / MaxViT [42]
	AFHQ [8]	StanfordDogs [9]	ResNeSt-101 [49]
MIRROR [3]	FFHQ [21]	VGGFace2 [4]	ResNet-50 [15]

Table 2. **Setups of our comprehensive experiments.** We follow the exact setups in the previous MI attacks. Beside the standard MI setups on GMI [51]/KEDMI [6] on VGG16, and VMI [43] on Resnet-34, we also evaluate our defense approach on current SOTA MI setups. Due to the need of intensive grid-search for hyper-paramters, it is very time consuming to expand the exisiting SOTA MI Defense [35] to these additional MI setups. In total, there are 20 MI setups spanning 7 MI attacks, 3 $\mathcal{D}_{pub}$, 4 $\mathcal{D}_{priv}$, 9 architectures of T. The experimental setups are described in more detail in the Supp.

[2, 22, 28]:

$$F = \mathbb{E}\left[-\frac{\partial^2}{\partial \theta_T^2}\mathcal{L}(X|\theta_T)\right] \tag{2}$$

Here, $\mathcal{L}$ is the loss function for a particular task. Specifically, we investigate FI on classification task and MI task. For classification, we follow Achille et al. [2] and Le et al. [27] to use cross entropy $\mathbb{E}\left[-\log p(y_i|x_i)\right]$ as $\mathcal{L}$ and validation set $\mathcal{D}_{priv}^{val} = \{(x_i, y_i)_{i=1}^{M}\}$ as X. For MI task, we propose to use the ℓ_2 distance between the feature representations of reconstructed images and the private images as $\mathcal{L}$:

$$\mathbb{E}\left[\left\|\Phi(\hat{x}_u^j) - \mathbb{E}\left[\Phi(x_{priv}^j)\right]\right\|_2\right] \tag{3}$$

Here, for a given input image, Φ computes the penultimate layer representation using the target model, and $\hat{x}_u^j$ is one of the MI reconstructed images for identity j, and $\mathbb{E}\left[\Phi(x_{priv}^j)\right]$ is the centroid feature of private images for identity j. Therefore, we use the distance between MI reconstructed image and private image of the same identity as the loss in FI analysis. The set of MI reconstructed images $\{\hat{x}_u^j\}_{j=1}^{J}$ for different identity is used as X. We explore different setups to compute $\mathcal{L}$, see Supp. In one setup, we perform FI analysis only at the last iteration (i.e., 3000, for the result in Fig. 1-II). As we are interested in FI at the layer level, we compute the average FI of all parameters within a layer. We use the main MI attack setup in Peng et al. [35], i.e., VGG16 with KEDMI [6] attack, for FI analysis.

Observation. The FI results in Fig. 1-II clearly suggest that the first few layers of a target model are important for MI task. Meanwhile, FI analysis suggests that the first few layers do not carry important information for a specific classification task. This observation is consistent with previous finding in work [48] suggesting that the earlier layers carry general features. The FI analysis justifies our design to prevent encoding of private information in the first few layers in order to degrade MI attacks, while keeping the impact on classification small. Overall, this leads to improved MI robustness. **Further results with different loss (ℓ_1 and LPIPS [50]) and different MI iterations can be found in Supp.**

4.3. Empirical Validation

As shown in Fig. 1-IV, we observe a significant improvement in MI robustness when reducing the number of parameters fine-tuned with $\mathcal{D}_{priv}$. However, the relationship between MI attack accuracy and natural accuracy is strongly correlated [51], which makes it unclear if the decrease in MI attack accuracy is due to the drop in natural accuracy.

In this section, we empirically investigate the hypothesis that *a model with fewer parameters encoding private information from $\mathcal{D}_{priv}$ has better MI robustness*. The empirical validation is reported in Fig. 1-III. Note that the number of parameters for the entire target model: $|\theta_C|$ = 16.8M for VGG16 with KEDMI [6] setup and $|\theta_C|$ = 11.7M for Resnet-18 with PPA [41] setup. The additional empirical validation for GMI can be found in the Supp. To separate the influence of model accuracy on MI attack accuracy, we perform PPA/KEDMI 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.

The results clearly show that fine-tuning fewer parameters on $\mathcal{D}_{priv}$ enhances MI robustness compared with fine-tuning all parameters on $\mathcal{D}_{priv}$, regardless of the effect on natural accuracy. For instance, in the KEDMI setup, with a comparable natural accuracy of 83%, fine-tuning only $|\theta_C|$ = 13.9M reduces a third attack accuracy compared to fine-tuning $|\theta_C|$ = 16.8M. The result in the PPA setup is even more supportive, where with a natural accuracy of around 91%, fine-tuning $|\theta_C|$ = 8.9M reduces the attack accuracy to 22.36% from 91.7% in $|\theta_C|$ = 11.7M.

Across all configurations, we observe that the fewer parameters fine-tuned on $\mathcal{D}_{priv}$, the more robust the model. However, it is important to note that if the number of fine-tuned parameters on $\mathcal{D}_{priv}$ is insufficient, such as $|\theta_C|$ = 9.1M for KEDMI setup, the model's natural accuracy may drop drastically, rendering it unusable. Overall, our experiments strongly suggest that **better MI robustness can be achieved by reducing the number of parameters fine-tuned on $\mathcal{D}_{priv}$.**

4.4. Comparison with SOTA MI Defense

In this section, we compare our proposed TL-DMI defense with current existing MI defenses [35, 44]. *For a fair comparison, we strictly follow the setups in SOTA MI defense [35].* Specifically, we first present the MI robustness comparison against KEDMI/GMI in Fig. 1-IV. MID [44] improves MI robustness by penalizing the mutual information between inputs and outputs during the training process, which is intractable in continuous and high-dimensional settings, making MID resort to mutual information approximations rather than actual quantity [35]. In general, MID is outperformed by more recent defense BiDO [35].

Our proposed TL-DMI is simple yet effective, achieving outstanding MI robustness as shown in Fig. 1-IV. We are the first to explore MI defense beyond the regularization perspective. TL-DMI can be combined with SOTA MI defenses such as BiDO. When combining with TL-DMI, we strictly follow BiDO. The only difference is that BiDO is applied only to the unfrozen layers in the fine-tuning stage. The results in Fig. 1-IV show that the trade-off between utility and robustness is much improved when we combine two approaches. Also, TL-DMI helps restore the utility degraded by BiDO, rendering a much more MI robust model (reducing MI attack accuracy by 27.36% from 46.23% to 18.87%) while improving model utility (increasing model accuracy by 1.8% from 80.35% to 82.15%).

In VMI setup presented in Tab. 3, MID [44] suffers when applied to VMI [43] due to the requirement of modifying the last layer of the network to implement the variational approximation of the mutual information [35]. Hence, we observe a significant drop in natural accuracy when applying MID [44] to VMI [43]. BiDO [35] partially addresses this problem and recovers natural accuracy better with comparable attack accuracy. Comapred to BiDO, TL-DMI updating $|\theta_C|$ = 21.14M (out of 21.5M parameters in totam) improves natural accuracy by around 1%-3% while achieving greater robustness by reducing attack accuracy by around 6%.

In another effort to comprehensively compare with the SOTA MI defense BiDO [35], we extend the evaluation to include additional SOTA MI attacks: LOMMA [32] and PPA [41]. Given the different setup of PPA compared to BiDO, we adapt BiDO to work with additional architectures of T, specifically ResNet-18/101 [15], tailored to the PPA attack. Note that these evaluations have not been explored yet in the MI literature [35, 44]. From the results in Tab. 3, a consistent trend is that all defenses have suffered in natural accuracy, but TL-DMI method has suffered the least in natural accuracy while reducing the most in attack accuracy. Consequently, TL-DMI achieves the best MI robustness trade-off, which can be quantified by Δ, which is the ratio of drop in attack accuracy to drop in natural accuracy (the larger is the ratio, the better is MI robustness). Additional comparison against BREPMI [19] can be found

Attack Method	T	Defense	Acc ⇑	AttAcc ⇓	Δ ⇑
VMI [43]	ResNet-34 [15]	No Def.	69.27	39.40	-
		BiDO	61.14	30.25	1.13
		TL-DMI	62.20	23.70	**2.22**
LOMMA [32]	VGG-16 [40]	No Def.	89.00	95.67	-
		BiDO	80.35	70.47	2.91
		TL-DMI	83.41	75.67	**3.58**
PPA [41]	ResNet-18 [15]	No Def.	94.22	88.46	-
		BiDO	91.33	76.56	4.12
		TL-DMI	91.12	22.36	**21.32**
	ResNet-101 [15]	No Def.	94.86	83.00	-
		BiDO	90.31	67.26	3.46
		TL-DMI	90.10	31.82	**10.75**

Table 3. The comparison between our proposed TL-DMI and SOTA MI denfense BiDO [35], where the Acc and AttAcc are given in %. Our evaluation covers a wide range of MI attack setups. We follow previous work for MI setups (see details in Tab. 2 and Supp.). To implement TL-DMI, we set $|\theta_C|$ = 21.14M/13.90M/8.90M/16.05M for T = ResNet-34/VGG-16/ResNet-18/ResNet-101, respectively. **MI robustness is quantified by the Δ, the ratio of drop in attack accuracy to drop in natural accuracy.** As shown in the results, our proposed TL-DMI significantly improves MI robustness comparing to BiDO.

in Supp. In conclusion, our proposed TL-DMI stands out as highly effective across a range of SOTA MI attacks [32, 41].

4.5. Extended MI Robustness Evaluation

Our proposed TL-DMI is simple, easy to implement, and less sensitive to hyperparameters than BiDO, which requires intensive grid search for hyperparameter. This significant advantage allows us to extend the scope of experimental setups for the MI defense to align with the remarkable increase in MI attack setups, which are not yet evaluated in previous MI defenses [35, 44].

Results on different $\mathcal{D}_{pub}$. We evaluate TL-DMI against KEDMI and GMI attacks on three architectures (VGG16, IR152, FaceNet64) with varying public datasets (CelebA, FFHQ), spanning 12 facial domain MI setups. These are standard setups in KEDMI/GMI, however, only 2 out of 12 setups examined in the current SOTA MI defense were presented in [35]. The results in Tab. 4 demonstrate that TL-DMI consistently achieves significantly more robust models across all setups while maintaining acceptable natural accuracy, with significant improvements in robustness across a wide range of attack scenarios (13.33%-42.60% for KEDMI, 11.14%-31.94% for GMI). *On average, TL-DMI significantly reduces the accuracy of MI attacks by more than a half.*

Results on SOTA high resolution MI attacks. Furthermore, we provide our defense results against SOTA High Resolution MI attacks, i.e., PPA [41] and MIRROR [3] in Tab. 3 and Tab. 5. To the best of our knowledge, our work is the first MI defense approach against such high resolution MI attack. The results are very encouraging. We observe only a small reduction in natural accuracy, while the attack accuracy experiences a significant drop thanks to our defense TL-DMI.

Results on different architectures of T. Unlike BiDO, TL-DMI does not require an intensive grid search for hyperparameter selection for a specific architecture. Therefore, TL-DMI offers high practicality and is readily applicable to a range of architectures, whereas existing state-of-the-art MI defenses lack this advantage [35]. We conduct evaluations on a range of architectures, including residual-based networks such as ResNet-18/50/101, ResNeSt-101, IR152, as well as the more recent MaxViT architecture [42]. Across all these experiments in Tab. 5 and Tab. 3, TL-DMI consistently demonstrate superior performance, highlighting its effectiveness and robustness across various architectures.

Result on different $\mathcal{D}_{priv}$. Regarding private dataset $\mathcal{D}_{priv}$, in addition to CelebA, which is standard for MI research, and other large-scale facial datasets including Facescrub [31] and VGGFace2 [4], our experiments go beyond these datasets by studying the animal domain, i.e., Stanford Dogs dataset [9]. The result is illustrated in Tab. 5. Via our comprehensive evaluation, we find that our approach consistently demonstrates its efficacy across various datasets, regardless multiple factors such as the number of training/attack classes or the specific domain under consideration. This versatility highlights the robustness and adaptability of our defense TL-DMI across a wide range of scenarios.

Overall, all these extensive results consistently support that our method is effective in defending against advanced MI attacks. Our approach is simple and can be easily applied, with minimal changes to the original training of target classifier T. **Additional results and analysis are included in the Supp.**

5. Conclusion

In this paper, we propose a simple and highly effective Transfer Learning-based Defense against Model Inversion (TL-DMI). Our method is a major departure from existing MI defense based on dependency minimization regularization. Our main idea is to leverage TL to limit the number of layers encoding private data information, thereby degrading the performance of MI attacks. To justify our method, we conduct the first study to analyze layer importance for MI task via Fisher Information. Our analysis results suggest that the first few layers are important for MI, justifying our design to prevent private information encoded in the first few layers. Our defense TL-DMI is remarkably simple to implement. Through extensive experiments, we demonstrate SOTA effectiveness of TL-DMI across 20 MI setups spanning 9 architectures, 4 private datasets D_{priv}, and 7 MI attacks.

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 ⇑
KEDMI	CelebA	CelebA	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	90.87 ± 2.71	99.33 ± 0.75	1168
					TL-DMI	13.9/16.8	83.41	**51.67 ± 3.93**	**80.33 ± 2.91**	**1410**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	94.07 ± 1.82	99.67 ± 0.63	1071
					TL-DMI	17.8/62.6	86.70	**64.60 ± 4.93**	**87.67 ± 2.73**	**1333**
				FaceNet64	No Def.	35.4/35.4	88.50	86.73 ± 2.85	98.33 ± 1.49	1194
					TL-DMI	34.4/35.4	83.41	**73.40 ± 4.10**	**91.67 ± 1.92**	**1265**
	CelebA	FFHQ	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	55.60 ± 3.75	84.67 ± 2.85	1407
					TL-DMI	13.9/16.8	83.41	**34.53 ± 3.43**	**65.33 ± 3.36**	**1554**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	70.27 ± 3.40	89.33 ± 2.14	1285
					TL-DMI	17.8/62.6	86.70	**46.53 ± 4.58**	**72.67 ± 3.16**	**1454**
				FaceNet64	No Def.	35.4/35.4	88.50	57.87 ± 4.70	82.00 ± 3.45	1409
					TL-DMI	34.4/35.4	83.41	**15.27 ± 4.09**	**31.00 ± 4.24**	**1751**
GMI	CelebA	CelebA	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	30.20 ± 5.26	55.00 ± 5.95	1600
					TL-DMI	13.9/16.8	83.41	**7.80 ± 3.36**	**23.33 ± 4.60**	**1845**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	40.87 ± 4.76	66.67 ± 5.76	1516
					TL-DMI	17.8/62.6	86.70	**8.93 ± 3.73**	**22.67 ± 5.21**	**1819**
				FaceNet64	No Def.	35.4/35.4	88.50	26.87 ± 3.75	49.00 ± 6.05	1643
					TL-DMI	34.4/35.4	83.61	**15.73 ± 4.58**	**33.00 ± 6.28**	**1752**
	CelebA	FFHQ	ImageNet1K	VGG16	No Def.	16.8/16.8	89.00	13.60 ± 4.43	32.00 ± 4.92	1725
					TL-DMI	13.9/16.8	83.41	**4.27 ± 2.56**	**12.33 ± 3.44**	**1919**
			MS-CelebA-1M	IR152	No Def.	62.6/62.6	93.52	24.27 ± 4.24	45.67 ± 6.71	1617
					TL-DMI	17.8/62.6	86.70	**6.13 ± 3.11**	**15.00 ± 4.98**	**1877**
				FaceNet64	No Def.	35.4/35.4	88.50	13.13 ± 4.96	30.33 ± 5.40	1746
					TL-DMI	34.4/35.4	83.61	**2.60 ± 1.49**	**8.67 ± 3.64**	**2009**

Table 4. Our evaluation covers multiple MI attack setups, target models, and public, private and pre-trained datasets. Here, the results are given in %. Specifically, we reports the MI defense results against different MI attack methods (KEDMI and GMI), as well as using different public datasets $\mathcal{D}_{pub}$ (CelebA and FFHQ), and pre-trained datasets $\mathcal{D}_{pretrain}$ (Imagenet1K and MS-CelebA-1M), for several target model T: VGG16, IR152, FaceNet64.

Attack Method	$\mathcal{D}_{priv}$	T	Defense	Acc ⇑	AttAcc ⇓	δ_{Eval} ⇑	$\delta_{FaceNet}$ ⇑	ℓ_2 Dist ⇑	FID ⇑
PPA	FaceScrub	MaxViT	No Def.	96.57	79.63	128.46	0.7775	-	50.37
			TL-DMI	93.01	**21.17**	**168.85**	**1.0199**	**-**	**55.50**
	Stanford Dogs	ResNeSt-101	No Def.	75.07	91.90	62.56	-	-	33.69
			TL-DMI	79.54	**60.88**	**83.57**	**-**	**-**	**46.01**
MIRROR	VGGFace2	ResNet-50	No Def.	99.44	84.00	-	-	602.41	-
			TL-DMI	99.40	**50.00**	**-**	**-**	**650.28**	**-**

Table 5. The defense results for SOTA MI attacks on 224x224 images. We strictly follow experimental setups from PPA and MIRROR, presenting results for Acc and AttAcc in %. Additionally, we employ PPA-introduced metrics, $\delta_{FaceNet}$ and δ_{Eval}, alongside MIRROR-introduced metric l_2 Dist for the evaluation. Our proposed TL-DMI successfully defends against SOTA MI attacks on high resolution 224x224. To train our TL-DMI defense models, we set $|\theta_C|$ = 18.3M/27.9M/32.9M for T = MaxViT/ResNeSt-101/ResNet-50, respectively.

Limitation. Following other MI attack and defense research [6, 32, 35, 44, 51], our current focus is on classification. However, our future work will extend to studying MI attacks and defenses for other machine learning tasks, such as object detection.

Ethical consideration. Our research on improving MI robustness addresses a significant ethical concern in modern data-driven machine learning: data privacy. Our study is based on publicly available standard data and does not involve the collection of sensitive information.

Acknowledgement. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); The Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This material is based on the research/work support in part by the Changi General Hospital and Singapore University of Technology and Design, under the HealthTech Innovation Fund (HTIF Award No. CGH-SUTD-2021-004).

References

[1] Milad Abdollahzadeh, Touba Malekzadeh, Christopher TH Teo, Keshigeyan Chandrasegaran, Guimeng Liu, and Ngai-Man Cheung. A survey on generative modeling with limited data, few shots, and zero shot. *arXiv preprint arXiv:2307.14397*, 2023. 1

[2] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019. 5

[3] 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. 3, 4, 5, 7

[4] 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. 5, 7

[5] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe. End-to-end multi-speaker speech recognition with transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6134–6138. IEEE, 2020. 1

[6] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16178–16187, 2021. 1, 3, 4, 5, 6, 8

[7] 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. 5

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 5

[9] E Dataset. Novel datasets for fine-grained image categorization. In *First Workshop on Fine Grained Visual Categorization, CVPR. Citeseer. Citeseer*. Citeseer, 2011. 5, 7

[10] 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. 4

[11] Jonas Dippel, Steffen Vogler, and Johannes Höhne. Towards fine-grained visual representations by combining contrastive learning with image reconstruction and attention-weighted pooling. *arXiv preprint arXiv:2104.04323*, 2021. 1

[12] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 58–68. Springer, 2021. 1

[13] 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. 1

[14] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6172, 2020. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 5, 6, 7

[16] Xinlei He, Hongbin Liu, Neil Zhenqiang Gong, and Yang Zhang. Semi-leak: Membership inference attacks against semi-supervised learning. In *European Conference on Computer Vision*, pages 365–381. Springer, 2022. 3

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 1

[18] Junguang Jiang, Yang Shu, Jianmin Wang, and Mingsheng Long. Transferability in deep learning: A survey. *Journal of Machine Learning Research*, 2022. 3, 4

[19] 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. 5, 6

[20] Uday Kamath, John Liu, James Whitaker, Uday Kamath, John Liu, and James Whitaker. Transfer learning: Domain adaptation. *Deep learning for NLP and speech recognition*, pages 495–535, 2019. 4

[21] 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. 5

[22] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016. 3, 5

[23] Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881, 2023. 3

[24] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Large scale learning of general visual representations for transfer. *arXiv preprint arXiv:1912.11370*, 2(8), 2019. 4

[25] Gautam Krishna, Co Tran, Jianguo Yu, and Ahmed H Tewfik. Speech recognition with no speech or with noisy speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1090–1094. IEEE, 2019. 1

[26] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 4

[27] Cat P Le, Mohammadreza Soltani, Juncheng Dong, and Vahid Tarokh. Fisher task distance and its applications in transfer learning and neural architecture search. *arXiv preprint arXiv:2103.12827*, 2021. 5

[28] Yijun Li, Richard Zhang, Jingwan Cynthia Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *Advances in Neural Information Processing Systems*, 33:15885–15896, 2020. 3, 5

[29] 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. 4, 5

[30] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 1

[31] 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. 5, 7

[32] 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. 1, 3, 4, 5, 6, 7, 8

[33] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai man Cheung. Label-only model inversion attacks via knowledge transfer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[34] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 1, 3, 4

[35] 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. 1, 3, 4, 5, 6, 7, 8

[36] Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021. 3

[37] 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. 1

[38] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. Membership inference attacks are easier on difficult problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14820–14829, 2021. 3

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

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4, 5, 7

[41] 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. 3, 4, 5, 6, 7

[42] 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. 3, 5, 7

[43] Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. *Advances in Neural Information Processing Systems*, 34:9706–9719, 2021. 1, 5, 6, 7

[44] 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. 1, 3, 4, 5, 6, 7, 8

[45] Jiawei Yang, Hanbo Chen, Jiangpeng Yan, Xiaoyu Chen, and Jianhua Yao. Towards better understanding and better generalization of few-shot classification in histology images with contrastive learning. 2022. 1

[46] 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. 4

[47] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019. 4

[48] 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. 1, 2, 3, 6

[49] 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. 5

[50] 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. 6

[51] 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. 1, 3, 4, 5, 6, 8

[52] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 4