Decoupling Generalizability and Membership Privacy Risks in Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

A deep learning model usually has to sacrifice some utilities when it acquires some other abilities or characteristics. Privacy preservation has such trade-off relationships with utilities. The loss disparity between various defense approaches implies the potential to decouple generalizability and privacy risks to maximize privacy gain. In this paper, we identify that the model's generalization and privacy risks exist in different regions in deep neural network architectures. Based on the observations that we investigate, we propose *Privacy-Preserving Training Principle* (PPTP) to protect model components from privacy risks while minimizing the loss in generalizability. Through extensive evaluations, our approach shows significantly better maintenance in model generalizability while enhancing privacy preservation.

1 Introduction

A machine learning model acquires accurate recognition abilities by learning how to fit the training data points. However, even if a model shows a good performance for an objective of a given task, it may suffer from the risk of data privacy leakage, especially in privacy-sensitive applications or systems. The concern about this issue is also raised by several existing studies: Arpit et al. (2017); Chatterjee (2018) showed neural networks' potential privacy risks via fitting on random data. Shokri et al. (2017) showed that privacy risks are widely posed in neural networks by showing the possibility of black-box membership inference attacks (MIAs). Further studies Choquette-Choo et al. (2021); Del Grosso et al. (2022) showed that a model has significant behavioral differences between the training and testing data in various aspects, especially in robustness. Also, Stephenson et al. (2021) found that the model's memory of training data points becomes solid as training progresses. All of these prior observations and knowledge indicate that machine learning models have a strong memory for data, which enables the model to achieve near-perfect performance on the training dataset.

Since a well-trained model will retain a lot of data traces, deploying a model in a privacy-sensitive system requires particular caution. This risk occurs not only in classification models Shokri et al. (2017); Song & Mittal (2021); Choquette-Choo et al. (2021) but also in some other machine learning domains, such as generative models Chen et al. (2020) or transfer learning Zou et al. (2020); Wu et al. (2024), etc. Besides, models could leak privacy in various ways, e.g., membership inference attacks Shokri et al. (2017), model inversion attacks Fredrikson et al. (2015), and model extraction attacks Tramèr et al. (2016). This universality makes the output of any model pose the potential risk of privacy leakage. Therefore, understanding the sources of privacy risks and how to rectify them is the key to strengthening the model to be trustworthy.

In this paper, we study where and how layer-level privacy leakage occurs in neural network architectures. We empirically identify that the model's generalizability and privacy risk are separable. Then, we propose Privacy-Preserving Training Principle (PPTP) to minimize negative impacts on the model utility during training with privacy defense approaches. Here is a brief overview of our novel observations and contributions:

- We structurally and precisely investigate where and how the machine learning model produces privacy risks, and identify that ***privacy risks and generalizability occur in different regions in a model.***

- We propose a new cost-effective training paradigm for utility-privacy trade-offs. It ***decouples utility and privacy*** into two separate parts, enabling the model to outperform existing privacy defense approaches.
- Our approach bridges the gap between generalizability and privacy training, empowering the model to choose its proper training approaches (for both utility and privacy) based on its application contexts and helping the model overcome the limitations of the existing privacy defense approaches.

2 Related Work

There have been various studies to prevent the ML model from privacy risks. DP-SGD Abadi et al. (2016) tried to train the model with a less fitting degree via additional noises in the optimizer. Nasr et al. (2018) developed an adversarial mechanism to help the model obtain aligned predictions. However, both require significantly increasing training costs due to their limitations in parallel computations. Jia et al. (2019); Yang et al. (2023) proposed decorators to reproduce better-aligned predictions without retraining the model. Shejwalkar & Houmansadr (2021) tried to mitigate the privacy risk via knowledge distillation. Tang et al. (2022) further improved the utilization rate of training data via ensemble-based knowledge distillation. Li et al. (2021) attempted the alignment between training and non-training accuracy during training. Chen et al. (2022) proposed an efficient training paradigm with effective privacy-preserving ability. Fang & Kim (2024a;b) discussed the privacy issues in the bottleneck layer. Liu et al. (2024) observed the convexity of loss functions is a factor of privacy leakage and tried to mitigate it with a concave term. Although the conclusions are not entirely consistent, Wang et al. (2021); Yuan & Zhang (2022) explored the impact of network pruning on privacy risks. Kaya & Dumitras (2021); Yu et al. (2021) explored which data augmentation techniques are beneficial to privacy. Tan et al. (2023) pointed out that higher model dimensions are possibly more privacy-risky. Li et al. (2024) tried to focus on the most privacy-risky data points. Carlini et al. (2022b) found that a model always experiences varying degrees for different data points in traditional training settings. Zhang et al. (2024) found some components lead to machine learning models at severe privacy risks.

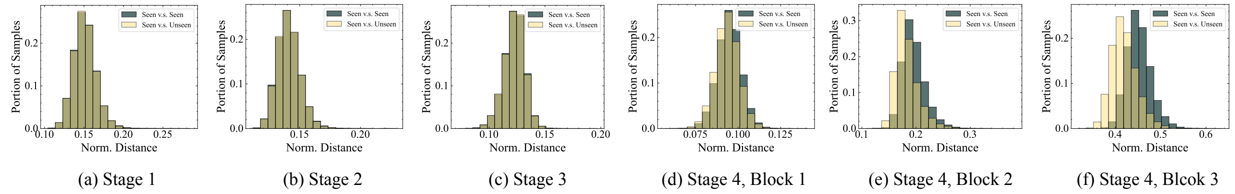
Despite considerable progress, it still remains a work in progress. Like obtaining other characteristics (e.g., Robustness Goodfellow et al. (2015), Fairness Mehrabi et al. (2021), and Data Unlearning Bourtoule et al. (2021)), models always inevitably have to pay the price of utility for privacy with current approaches. Sometimes, this expense is even too high for the model to maintain reasonable performance due to the inherent characteristics of the dataset Carlini et al. (2022a). We question the preconceived view kept by current privacy-preserving approaches that the model is a privacy-risky entity and explore where the model produces the privacy risks in the next section.

3 Does Prediction Disparity Exist Everywhere?

3.1 Why the Question Matters

Recent non-decorator defense approaches Abadi et al. (2016); Nasr et al. (2018); Shejwalkar & Houmansadr (2021); Chen et al. (2022); Tang et al. (2022) usually train a model from scratch while decorator approaches Jia et al. (2019); Yang et al. (2023) usually add extra filters to the model externally rather than training the model itself. A direct advantage of training from scratch is that it is implementation-friendly. However, training the whole model could lead to an unnecessary utility loss as we see a discrepancy between adversarial training on the whole model Shafahi et al. (2019) and training the last layer only Kirichenko et al. (2023). Then, it gives us a question: **Does prediction disparity exist everywhere**? In this section, we discuss the correlation between privacy-risk and other attributes of machine learning models, such as feature map size, channel size, and depth.

It is known that the model's generalizability is affected by the depth Baldock et al. (2021). That is, a model gradually obtains generalizability layer by layer while learning samples with various difficulty levels. If privacy risk accompanies the learning of different difficulty-level samples, the privacy risk should exist in each layer and "*gradually*" show more and more prediction disparity. Meanwhile, Kirichenko et al. (2023) showed that spurious correlations can be mitigated by retraining only the last layer, hinting that different



(a) Stage 1 (b) Stage 2 (c) Stage 3 (d) Stage 4, Block 1 (e) Stage 4, Block 2 (f) Stage 4, Blcok 3

Figure 1: The sample-level feature map differences (norm. distance on x-axis). No disparity is observed in Stage 1–3, whereas gradually increasing disparity is observed within Stage 4 (ResNet152, TinyImageNet, data augmented)

layers may not exhibit the same characteristics with regard to generalizability. Such recent insights challenge us to locate where a disparity between generalizability and privacy risks starts occurring.

3.2 Sample-Level Measurement Design

Unlike the logits produced by the classification layer, there is no direct way to know what a confident feature map should look like. Therefore, we develop a method to examine distribution disparity. Let D_{all} denote the entire training set. We split D_{all} into two halves: D_{h1} and D_{h2}. Then, we train two models, M_{all} and M_{h1}, with the same architecture and configurations with training sets D_{all} and D_{h1}, respectively. Therefore, every data point $x_1 \in D_{h1}$ is seen data for both two models, while $x_2 \in D_{h2}$ is seen data for only M_{all} but unseen data for M_{h1}.

Then, we compute the feature map differences using Euclidean distance $\texttt{Dist}(p, q) = \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2/d}$, where d is the number of dimensions of the feature maps and $1/\sqrt{d}$ is an extra normalized item, while p and q denote the feature maps produced by M_{all} and M_{h1} on the same input, respectively. If a layer does not produce prediction disparity, then the feature map distance distributions on D_{h1} and D_{h2} should be equivalent, and vice versa.

3.3 Empirical Verification

First of all, we need to check if the disparity happens in all layers or not. Fig. 1 plots sample-level feature map differences of seen and unseen data (norm. distance on x-axis) where we find that the disparity ***abruptly*** starts in stage 4. In other words, there is only a small portion (the 4th stage contains only 3 blocks while the entire model contains 33 blocks) of ResNet152 that leaks privacy, identifying that the model is not globally privacy-risky, but *regionally.* However, what conditions cause disparity is not clear yet. Hence, we explore different factors' impacts on disparity in the rest of this section. We explore these factors using three different architectures as shown in Fig. 2: VGG Simonyan & Zisserman (2015b), ResNet He et al. (2016), and Active Token Mixer (ATM) (which is a transformer-type architecture) Wei et al. (2023). In the three architectures, ResNet and ATM have residual connections, while VGG does not. ATM has self-attention-type computation modules, which differentiates it from VGG and ResNet.

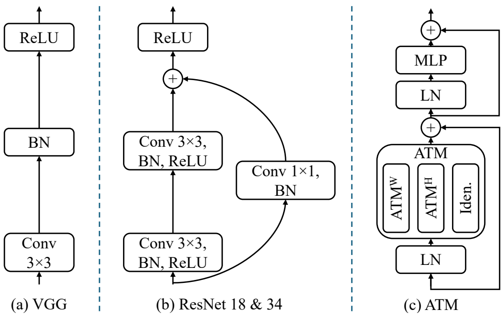


Figure 2: Overview of the three architectures' backbone modules.

Data Augmentation Kaya & Dumitras (2021); Yu et al. (2021) found some data augmentation techniques are beneficial to the model's privacy. Figure 3 shows that data augmentation does help mitigate disparity when it occurs (please compare (d) and (h)). However, whether with or without data augmentation, disparity happens in stage 4, which exhibits that data augmentation has no impact on "where" disparity happens.

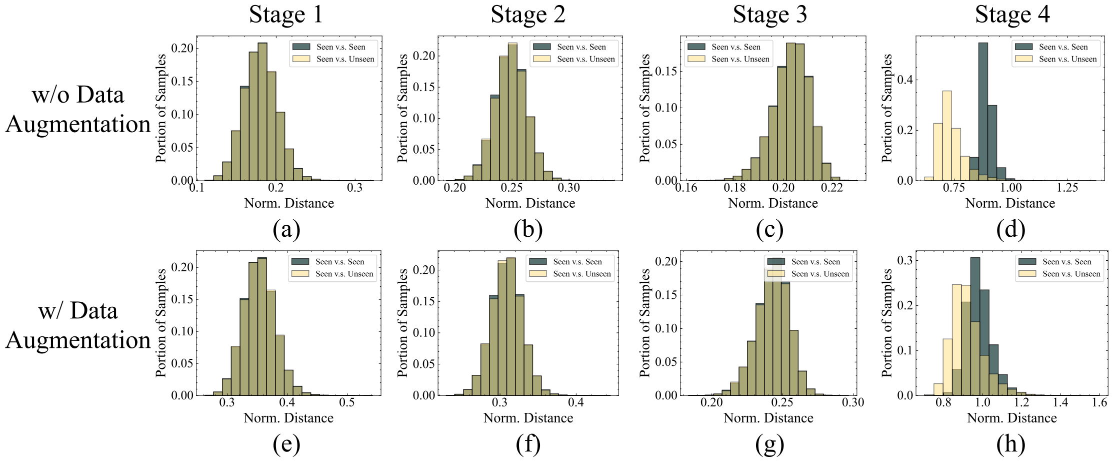


Figure 3: Comparison of models trained with and without data augmentation. (ResNet18, TinyImageNet)

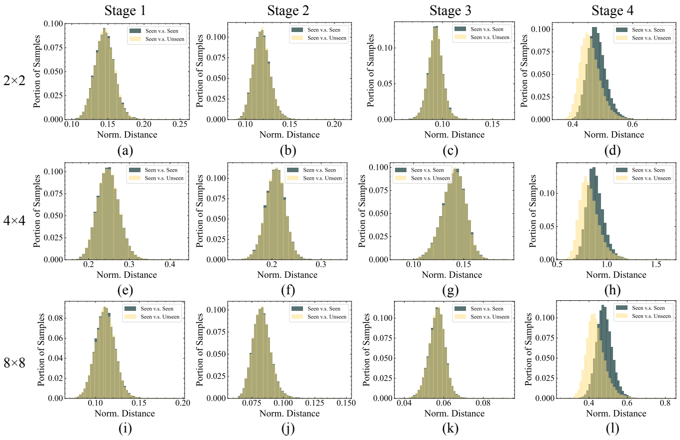


Figure 4: Comparison of ResNet18 with different feature map sizes in the 4th stage. (TinyImageNet, data augmented)

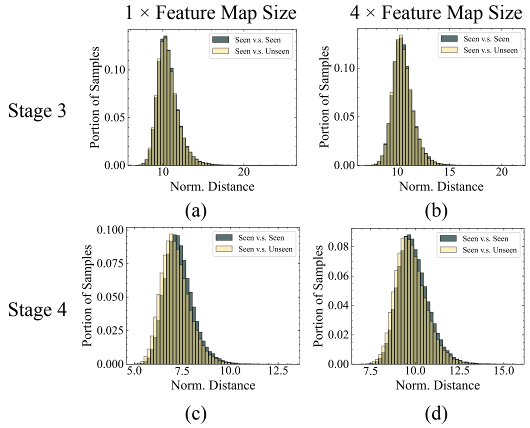


Figure 5: Comparison of ATM-XT in various channel sizes at the 3rd & 4th stage. (TinyImageNet, data augmented).

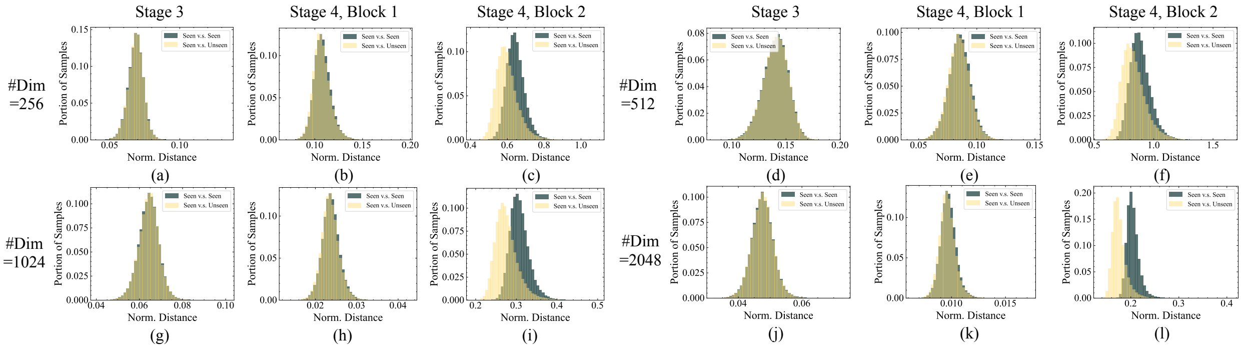


Figure 6: Comparison of ResNet18 in various channel sizes at the 4th stage. (TinyImageNet, data augmented)

Feature Map Size One significant change along with stages is the feature map size. After each downsampling layer at the head of each stage in the ResNet (similar to many widely used architectures such as CNN, ViT, Mixer, etc.), the feature map's width and height are halved. To study how generalizability

Table 1: The performance comparison among different feature map sizes at the 4th stage. (ResNet18, TinyImageNet, data augmented)

	Feature map size		
Accuracy (%)	2×2	4×4	8×8
Train	99.98	99.98	99.98
Test	52.32	54.58	61.53

Table 2: The performance comparison among different feature map sizes at the 4th stage. (ATM-XT, TinyImageNet, Data Augmentation)

Accuracy (%)	1× Feature Map Size	4× Feature Map Size
Train	99.97	99.98
Test	36.28	47.29

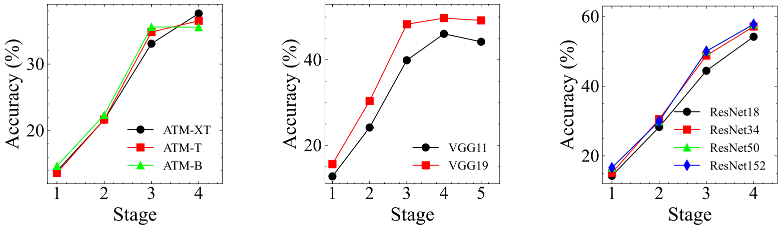
Table 3: The testing accuracy (%) comparison among different training data and different channel sizes at the 4th stage. (ResNet18, TinyImageNet, data augmented)

	Channel size			
Accuracy (%)	256	512	1024	2048
Test	52.52	54.58	53.99	55.03

and disparity change along with feature map size, the feature map size at the 4th stage of ResNet18 is enlarged and reduced (originally 4×4). By considering both Figure 4 and Table 1 together, we find that, while enlarging the feature map also has no impact on where the disparity happens, it enhances the model's generalizability while not significantly exacerbating the disparity. The same trends are also observed in ATMs (refer to Figure 5 and Table 2)

Channel Size One common design trend in deep learning in the recent decade is increasing channel size. As shown in Fig. 6 and Table 3, the disparity becomes more and more significant as channel size increases while changes in testing accuracy are not as significant as the changes by enlarged feature map size (refer to Table 1) at 4th stage. That is, too many channel sizes make the model prone to produce disparity, leading to more severe privacy risks. Therefore, it is not worth designing too many channels from the point of view of utility-privacy trade-offs.

Depth As observed above, since feature map size and channel size do not impact where privacy risk occurs, now we investigate the depth of the model. First, we use a support vector machine (SVM) Cortes & Vapnik (1995) to classify the features extracted in each stage (see Figure 7). We notice that some stages have limited (or even no) benefits to the model's generalizability. The testing accuracy in all models keeps consistently increasing in the early stages (stages 1, 2, and 3 in Figure 7). However, this growth rate starts to level off after stage 3 to various degrees - some of them slow down (ResNet), and others stagnate or even deteriorate (ATM and VGG). Please note that this stagnation of growth perfectly matches where the privacy risk occurs (e.g., Figure 8 & Figure 7(b)). It explains that when the model cannot learn sufficient features to be generalized enough as much as its capacity, it instead learns many ineffective features. These features cannot help generalizability but can help the model fit better on the train set. Therefore, these features must be privacy-risky since they are only valid on the train set.



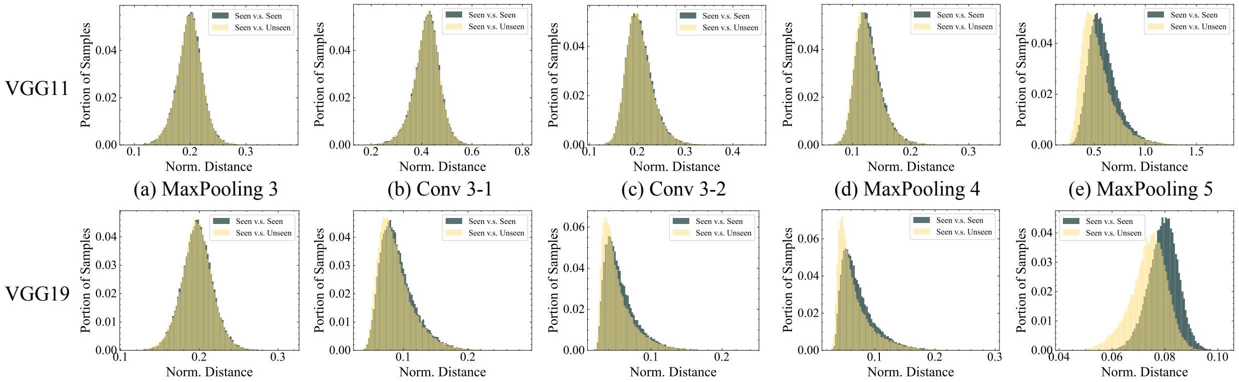
(a) ATM

(b) VGG

(c) ResNet

Figure 7: Testing accuracy changes along with stages using SVM. (TinyImageNet)

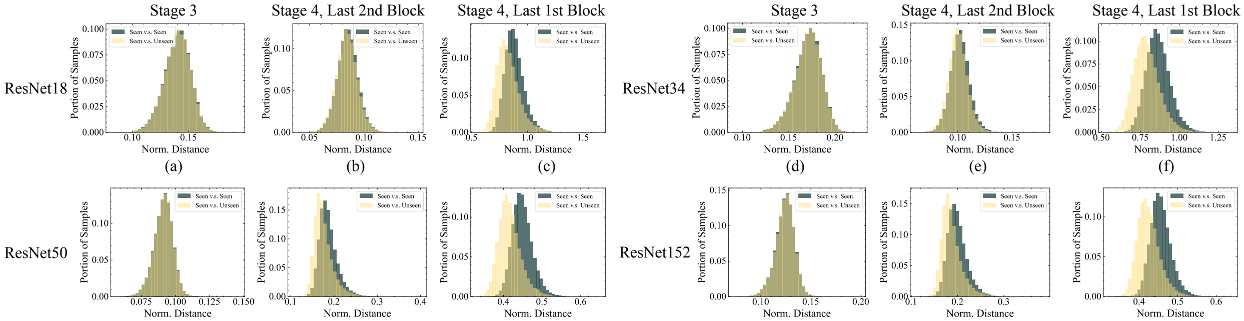
Besides, another notable point is that VGG shows different trends (privacy risks occur at different layers in VGG11&19, see Figure 8) from ATM's case (privacy risks always start at the 3rd stage, see Figure 10) and ResNet's case (privacy risks always start at the 4th stage, see Figure 9). That is because VGG11 can maintain the growing speed of generalizability between stages 3 and 4 while VGG19 cannot. In contrast, ResNet never shows privacy risks before the 4th stage even if most additional layers from ResNet18 to ResNet152 belong to the 3rd stage. This phenomenon further explains the relationship between generalized features and privacy.



(a) MaxPooling 3 (b) Conv 3-1 (c) Conv 3-2 (d) MaxPooling 4 (e) MaxPooling 5

(f) MaxPooling 3 (g) Conv 3-3 (h) Conv 3-4 (i) MaxPooling 4 (j) MaxPooling 5

Figure 8: The sample-level feature map differences measurement. The disparity is observed earlier in VGG19. (VGG11&19, TinyImageNet, data augmented)



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

Figure 9: Comparison of feature map differences among different ResNet depths.

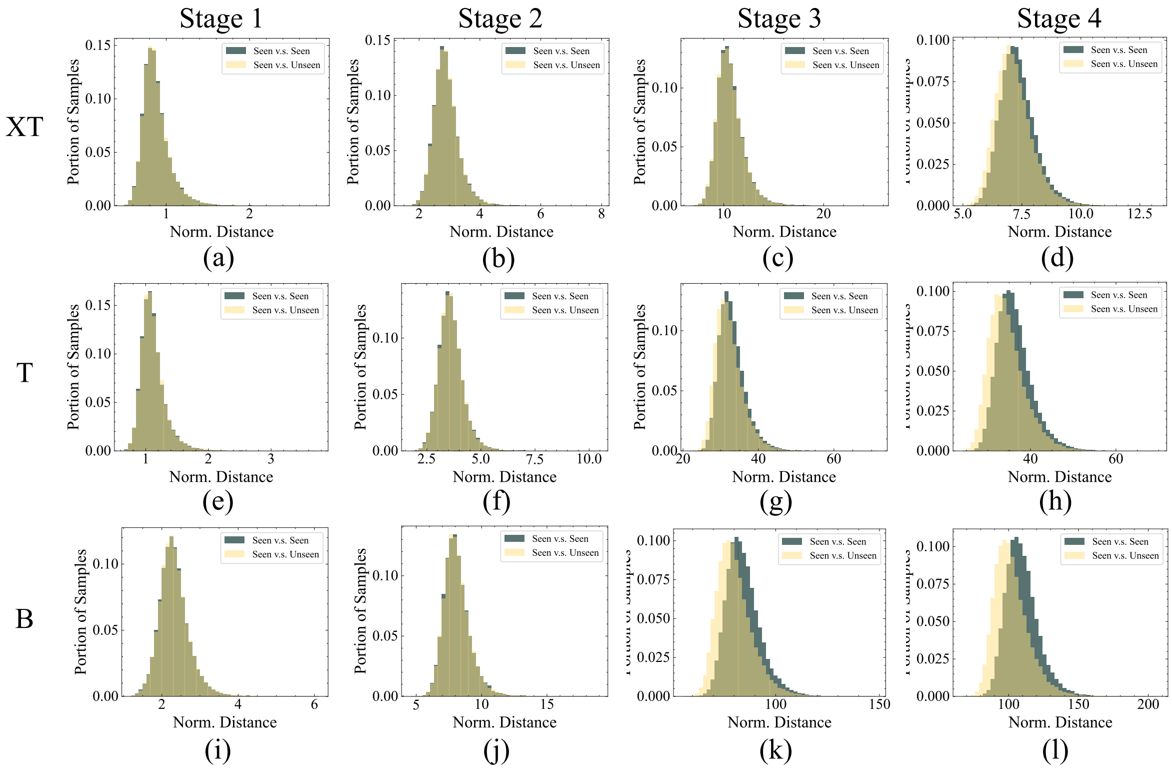
Does Prediction Disparity Exist Everywhere? The answer is no! In short, privacy risks only exist *in certain components without significant contributions to the model's generalizability.* Summarizing the observations above, we can draw the following insights of privacy and generalizability of an architecture:

- **Privacy risk and generalizability are separable**: in standard training procedures, privacy-risky features are learned together with generalized features due to imperfect or inadequate designs in various aspects, such as loss function, label, and model architecture). Fortunately, they exist in different regions of neural networks, identifying the feasibility of decoupling them.

- **Non- or less-generalized features lead to privacy risks**: Less-generalized features are usually learned in later layers. Although the increase in model computation capacity (i.e., increase in depth, channel size, and feature map size) brings the model effective improvements in generalizability, the surplus non- or less-generalized features put the model at higher privacy risks.

- **Privacy risk occurs in later stages**: The less-generalized features mainly exist in later stages, which is also consistent with the observation of Baldock et al. (2021). The more challenging to learn the feature is, the more possibly under-generalized the feature is. That is, privacy risk mostly exists in the later stage of the neural network.

These insights identify an issue in the current privacy defense approaches: they treat the model as a whole during training without distinguishing the privacy risks of various components. With no doubt, disturbing generalized privacy-safe features will lead to the unnecessary deterioration of the model's generalizability. Hence, it is necessary to develop a more considerate training paradigm with regard to privacy.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 10: Comparison of models in various depths. XT, T, and B denote ATM-XT, -T, and -B, respectively. (ATM, TinyImageNet)

4 Privacy-Preserving Training Principle

With the observations above in mind, we obtain the insight that a model does not need to update all weights to achieve better privacy-utility trade-offs. Instead, only privacy-risky layers can be retrained in a privacy-safe way. To achieve this, we propose a retraining approach, *Privacy-Preserving Training Principle* (PPTP), to enable a model to minimize utility loss while obtaining privacy. The approach is demonstrated in Alg. 1. In the algorithm, we first determine the privacy-risk and -safe layers. Then, *the weights of privacy-safe layers are frozen since they have learned generalized and privacy-safe features.* After that, the weights of the privacy-risky layers are rewound for retraining. Because general training approaches do not take into account privacy criteria, these layers need to be trained with a privacy-defending approach to be privacy-safe. With our approach, PPTP, the model can maintain highly generalized features while reducing privacy-risky features. We empirically show its effectiveness by comparing differences between training with and without our approach in the next section.

Algorithm 1 Privacy-Preserving Training Principle (PPTP)

```
Input: Training Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, Ordinarily Pre-Trained and Privacy-Risky Model $M_{pr}$ and its Parameters $\theta$, Privacy-Preserving Training Approach $f$, Training Epochs $E$, and Other Retraining Configurations $C$;
Output: Privacy-Safe Model $M_{ps}$;
Split the model with parameters $\theta$ into Privacy-Safe and -Risk Layers, denoting their parameters as $\theta_{ps}$ and $\theta_{pr}$, respectively.
Freeze Privacy-Safe Parameters $\theta_{ps}$
Rewind Privacy-Risky Parameters $\theta_{pr}$
for epoch in $\{1, 2, \cdots, E\}$ do
    Retrain the model with privacy-preserving training approach $f(\mathcal{D}, M_{pr}, C)$
end for
Return privacy-safe model $M_{ps}$
```

Table 4: Hyper-parameters searching space.

Hyper-Parameter	Adv-Reg	SELENA	RelaxLoss
α	$[1.0, 5.0]$	N/A	$[1.0, 3.0]$
k	3	25	N/A
l	N/A	$[3, 10]$	N/A

Table 5: The information about computation environment.

	OS	CPU	RAM	GPU	CUDA	Pytroch
Information	Ubuntu 22.04	96 cores	460 GB	V100	12.1	2.1

5 Experiments

5.1 Experimental Settings

We evaluate our approach and others on CIFAR-100 Krizhevsky et al. (2009) and TinyImageNet Le & Yang (2015). We evaluate our approaches on these two datasets to show the effectiveness on small and large datasets since extensive studies Chen et al. (2018); Yun et al. (2020) have demonstrated that methods that work well on small datasets may have insignificant improvement on larger datasets. Besides, the layers to be frozen in a model may vary over different datasets, especially with different input dimensions or information domains. In other words, the model may produce privacy risks earlier or later depending on the dataset. Data augmentation techniques, including random flipping & cropping Simonyan & Zisserman (2015a), are applied when training the model with a privacy-defending technique. As for privacy attacks, we evaluate our approach and others on correctness-based MIAs Yeom et al. (2020), confidence-based MIAs Yeom et al. (2018); Song et al. (2019); Song & Mittal (2021), entropy-based MIAs Shokri et al. (2017); Song & Mittal (2021), modified-entropy-based MIAs Song & Mittal (2021), and neural network based MIAs Shokri et al. (2017); Nasr et al. (2018). We include privacy training techniques such as adversarial regularization (`Adv-Reg`) Nasr et al. (2018), SELENA Tang et al. (2022), and relaxed loss (`RelaxLoss`) Chen et al. (2022). For each result, we execute three independent runs to ensure its stability. The hyper-parameters searching space of cited privacy defense approaches are shown in Table 4. Also, the main information of experimental software and hardware environment is presented in Table 7.

5.2 Results and Discussions

5.2.1 Ablation Study

To change the characteristics of a trained model, we can consider the two options, weight rewinding and fine-tuning. As seen in Algorithm 1, we choose rewinding and retraining to achieve more advanced utility-privacy trade-offs. Most privacy defense approaches, such as Nasr et al. (2018); Shejwalkar & Houmansadr (2021); Chen et al. (2022), are designed for training from scratch. Thus rewinding is more adequate for them. Besides, it is necessary to check if retraining privacy-risky layers can help the model achieve better utility-privacy trade-offs. To examine it, we show the performance comparison between training from scratch, retraining privacy-risky layers, and retraining the classification layers using RelaxLoss. As seen in Figure 11, retraining with the last layer only, regardless of whether through fine-tuning or weights rewinding, cannot effectively mitigate the model's privacy risks, while retraining multiple privacy-risky layers helps the model mitigate the privacy risks successfully.

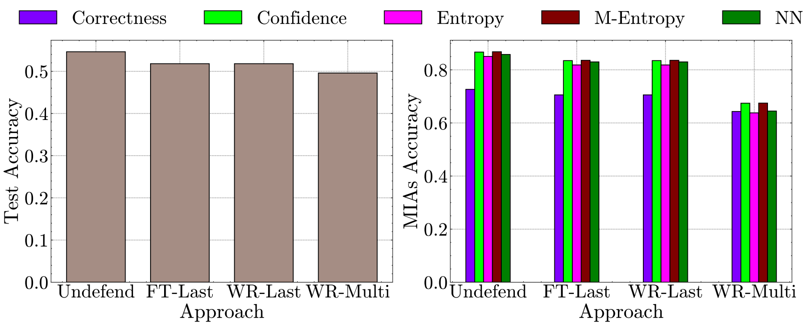


Figure 11: Comparison of retraining the model with different designs (TinyImageNet, ResNet18).

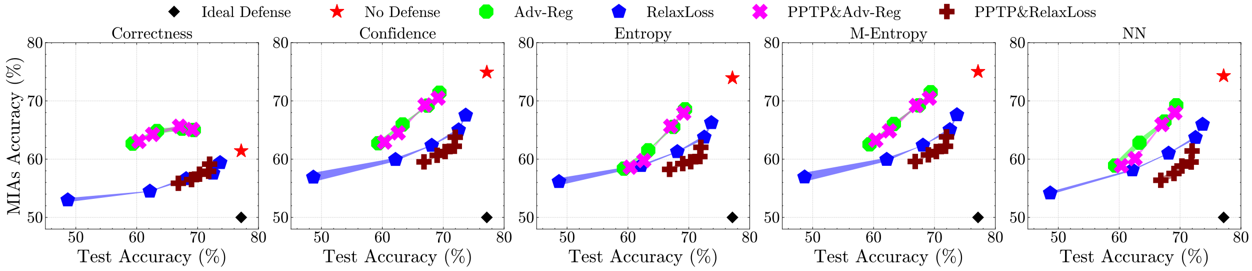


Figure 12: Comparisons with existing privacy-preserving techniques (CIFAR100, ResNet18).

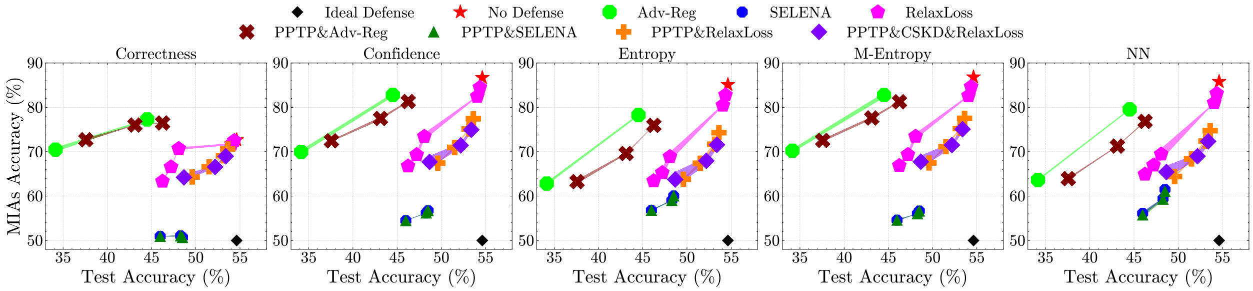


Figure 13: Comparisons with existing privacy-preserving techniques (TinyImageNet, ResNet18).

5.2.2 Comparisons with Other Approaches

In CIFAR-100, we evaluate our approach with Adv-Reg and RelaxLoss. As shown in Figure 12, both RelaxLoss and Adv-Reg show improvement when they are used to retrain the model with PPTP. However, the performance of correctness-based MIAs does not show significant improvement. PPTP retains the partially generalized features learned by CE to help the model maintain its utility, whereas it does not benefit the correctness alignment on train and test sets. Fortunately, however, this does not mean that PPTP cannot work with MIAs based on robustness or fairness. For instance, robustness issues of neural networks mainly exist in the final layer Kirichenko et al. (2023), meaning that the adversarial disparity can be also mitigated in the later layers to defend the robustness-based MIAs such as Del Grosso et al. (2022). We do not report the results of SELENA on this dataset since gradient explosions frequently occur on SELENA in the distillation stage (this phenomenon also occurs on TinyImageNet, although less frequently.)

In TinyImageNet, we evaluate our approach with Adv-Reg, SELENA, and RelaxLoss. Compared with results on CIFAR-100, the trends on TinyImageNet (Fig. 13) vary a lot. The Adv-Reg and RelaxLoss show more significant improvement once PPTP is applied. The results identify that our approach is effective in maintaining good performance on larger datasets. An important factor is that we freeze much more weights in TinyImageNet than in CIFAR-100 because the disparity occurs in different layers (see Table 6). This also shows the potential of the model compression methods Zhang et al. (2019; 2022), which can help the model achieve better utility performance at earlier stages, to be applied for privacy preservation. In contrast, the improvement in SELENA is very slight. We factor in

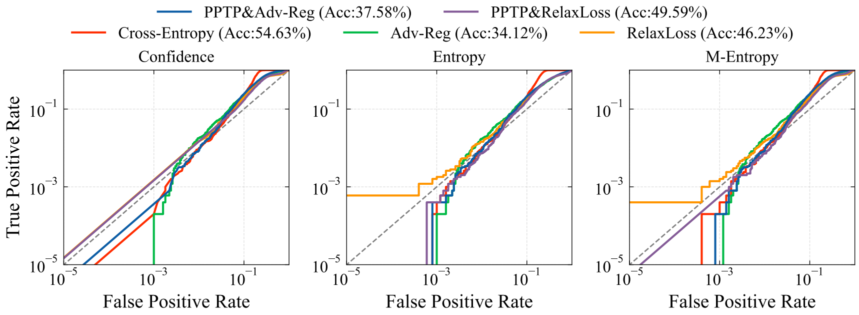


Figure 14: AUC-ROC curve comparison of defense approaches with and without PPTP under various MIAs.

Table 6: Weight-freezing stopping layers

Model	CIFAR-100	TinyImageNet
ResNet18	conv4_x (stage 3)	conv5_x (stage 4)

Table 7: The performance comparison of different training data and training techniques. (ResNet18, TinyImageNet, data augmented)

Accuracy (%)	CE & Full Data	CE & Half Data	RxL & Full Data
Train	99.98	99.99	79.68
Test	54.58	44.07	47.57

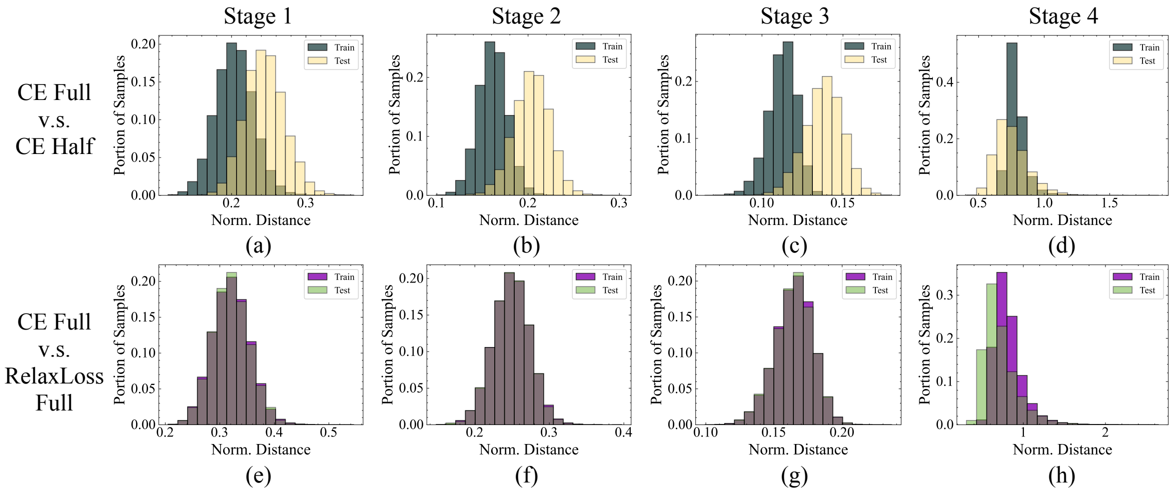


Figure 16: The distribution differences between training full data with CE, half data with CE, and full data with RelaxLoss.

that data augmentation is disabled in the distillation phase - if data augmentation was applied in this phase, the training time cost could have been over ten times greater than the current one. Additionally, it exhibits that our approach decouples the utility and privacy training. To demonstrate this characteristic, we apply class-wise self-knowledge distillation (CSKD) Yun et al. (2020) to pre-train the model. Compared with privacy training from scratch, pre-training with CSKD also shows improvements in utility at the same privacy level. Finally, to further exhibit privacy improvement using our approach, we evaluate Adv-Reg and RelaxLoss with and without PPTP and plot the AUC-ROC curve Carlini et al. (2022a) for TinyImageNet. As shown in Figure 14, compared with original Adv-Reg and RelaxLoss, our approach achieves comparable or better privacy with higher testing accuracy, further identifying our approach's effectiveness.

5.2.3 Training Cost

The privacy training cost can be optimized via PPTP since PPTP freezes a large portion of the weights. We evaluate the GPU memory and time cost to show the efficiency benefit that PPTP enables through privacy training approaches. As shown in Figure 15, the CUDA memory and time costs, evaluated on NVIDIA Tesla V100, clearly decrease after PPTP is applied. In particular, the actual training time cost is much less than the original training approaches (w/o PPTP) since retraining partial weights requires fewer epochs (less than half epochs) to converge.

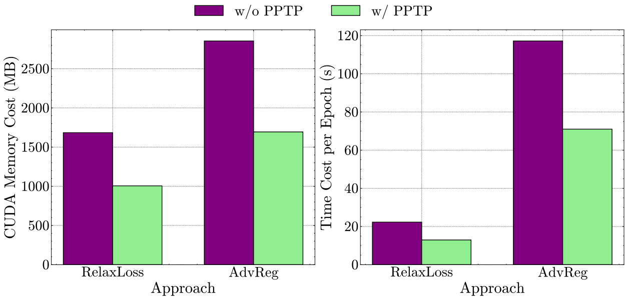


Figure 15: Our proposed Privacy-Preserving Training Principle (PPTP) clearly decreases training costs in memory and time. (CIFAR100, ResNet18).

5.2.4 Feature Map Differences

A significant factor that makes the model less generalized with a privacy-defending approach is that the feature representations in each model layer are quite different. As shown in Figure 16, the differences in feature maps between training with RelaxLoss and CE are more significant than training with different data (full vs. half). However, closer feature maps do not mean better generalizability (please refer to Table 7). These identify that our approach can preserve the generalized features learned by CE at privacy-safe layers and help the privacy-risky layers become more privacy-safe to mitigate privacy leakage.

6 Conclusion

In this paper, we observed that privacy vulnerability occurs at a portion of layers rather than the entire network. We underscore that the generalizability and privacy risks are decomposable since the well-generalized features and privacy-risky features exist in the different regions of the model. With this insight, we proposed Privacy-Preserving Training Principle (PPTP) to preserve generalizability along with privacy training. Through extensive empirical results, we showed that our approach enhances privacy with the proposed efficient training while not losing generalizability.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/arpit17a.html`.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10876–10889. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/5a4b25aaed25c2ee1b74de72dc03c14e-Paper.pdf`.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022a.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL `https://openreview.net/forum?id=ErUlLrGaVEU`.

Satrajit Chatterjee. Learning and memorization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 755–763. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/chatterjee18a.html`.

Binghui Chen, Weihong Deng, and Haifeng Shen. Virtual class enhanced discriminative embedding learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran

Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d79aac075930c83c2f1e369a511148fe-Paper.pdf.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *ACM Conference on Computer and Communications Security (CCS)*, 2020.

Dingfan Chen, Ning Yu, and Mario Fritz. Relaxloss: Defending membership inference attacks without losing utility. In *International Conference on Learning Representations*, 2022.

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

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Ganesh Del Grosso, Hamid Jalalzai, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. Leveraging adversarial examples to quantify membership information leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10399–10409, 2022.

Xingli Fang and Jung-Eun Kim. Center-based relaxed learning against membership inference attacks. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024a. URL https://openreview.net/forum?id=unlWrunFjg.

Xingli Fang and Jung-Eun Kim. Representation magnitude has a liability to privacy vulnerability, 2024b. URL https://arxiv.org/abs/2407.16164.

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

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*, pp. 770–778, 2016.

Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 259–274, 2019.

Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5345–5355. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kaya21a.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge, 2015.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pp. 5–16, 2021.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. {MIST}: Defending against membership inference attacks through {Membership-Invariant} subspace training. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2387–2404, 2024.

Zhenlong Liu, Lei Feng, Huiping Zhuang, Xiaofeng Cao, and Hongxin Wei. Mitigating privacy risk in membership inference by convex-concave loss. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30998–31014. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/liu24q.html`.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL `https://doi.org/10.1145/3457607`.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pp. 634–646, 2018.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/7503cfacd12053d309b6bed5c89de212-Paper.pdf`.

Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9549–9557, 2021.

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)*, pp. 3–18. IEEE, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015a.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015b.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2615–2632, 2021.

Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pp. 241–257, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367479. doi: 10.1145/3319535.3354211. URL `https://doi.org/10.1145/3319535.3354211`.

Cory Stephenson, suchismita padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=V8jrrnwGbuc`.

Jasper Tan, Daniel LeJeune, Blake Mason, Hamid Javadi, and Richard G. Baraniuk. A blessing of dimensionality in membership inference through regularization. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10968–10993. PMLR, 25–27 Apr 2023.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In *31st USENIX Security Symposium (USENIX Security)*, 2022.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Yijue Wang, Chenghong Wang, Zigeng Wang, Shanglin Zhou, Hang Liu, Jinbo Bi, Caiwen Ding, and Sanguthevar Rajasekaran. Against membership inference attack: Pruning is all you need. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3141–3147, 2021.

Guoqiang Wei, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. Active token mixer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2759–2767, 2023.

Cong Wu, Jing Chen, Qianru Fang, Kun He, Ziming Zhao, Hao Ren, Guowen Xu, Yang Liu, and Yang Xiang. Rethinking membership inference attacks against transfer learning. *IEEE Transactions on Information Forensics and Security*, 19:6441–6454, 2024. doi: 10.1109/TIFS.2024.3413592.

Ziqi Yang, Lijin Wang, Da Yang, Jie Wan, Ziming Zhao, Ee-Chien Chang, Fan Zhang, and Kui Ren. Purifier: Defending data inference attacks via transforming confidence scores. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10871–10879, 2023.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Samuel Yeom, Irene Giacomelli, Alan Menaged, Matt Fredrikson, and Somesh Jha. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 28(1):35–70, 2020.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. How does data augmentation affect privacy in machine learning? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10746–10753, 2021.

Xiaoyong Yuan and Lan Zhang. Membership inference attacks and defenses in neural network pruning. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 4561–4578, 2022.

Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13876–13885, 2020.

Guangsheng Zhang, Bo Liu, Huan Tian, Tianqing Zhu, Ming Ding, and Wanlei Zhou. How does a deep learning model architecture impact its privacy? a comprehensive study of privacy attacks on {CNNs} and transformers. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 6795–6812, 2024.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019.

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2022.

Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning, 2020. URL `https://arxiv.org/abs/2009.04872`.