Rethinking Membership Inference Attacks Against Transfer Learning

Cong Wu, Jing Chen, Qianru Fang, Kun He, *Member, IEEE*, Ziming Zhao, Hao Ren, Guowen Xu, Yang Liu, *Senior Member, IEEE*, and Yang Xiang, *Fellow, IEEE*

***Abstract*— Transfer learning, successful in knowledge translation across related tasks, faces a substantial privacy threat from membership inference attacks (MIAs). These attacks, despite posing significant risk to ML model's training data, remain limited-explored in transfer learning. The interaction between teacher and student models in transfer learning has not been thoroughly explored in MIAs, potentially resulting in an under-examined aspect of privacy vulnerabilities within transfer learning. In this paper, we propose a new MIA vector against transfer learning, to determine whether a specific data point was used to train the teacher model while only accessing the student model in a white-box setting. Our method delves into the intricate relationship between teacher and student models, analyzing the discrepancies in hidden layer representations between the student model and its shadow counterpart. These identified differences are then adeptly utilized to refine the shadow model's training process and to inform membership inference decisions effectively. Our method, evaluated across four datasets in diverse transfer learning tasks, reveals that even when an attacker only has access to the student model, the teacher model's training data remains susceptible to MIAs. We believe our work unveils the unexplored risk of membership inference in transfer learning.**

***Index Terms*— Membership inference attack, transfer learning.**

I. Introduction

TRANSFER learning has witnessed a marked increase in adoption recently, primarily due to its proficiency in leveraging knowledge from one domain to enhance performance in another, closely related domain [1], [2]. This method enables the application of complex models developed with extensive data, such as those by large corporations or hospitals, to be adapted for use in smaller entities like startups or smaller hospitals, thereby bypassing the need for substantial labeled data. However, this widespread adoption also raises concerns regarding the potential for privacy breaches, particularly through MIAs [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], as sensitive information, including biometric and healthcare data, is increasingly shared among different organizations [17]. Consequently, while transfer learning offers numerous advantages, such as customizing models to specific tasks without extensive data, it also necessitates stringent scrutiny to prevent potential privacy violations.

In MIAs, adversaries exploit access to a model's parameters or its output to ascertain if a specific data point was used during the model's training [3], [7]. They typically achieve this by analyzing how the model behaves with a given data point compared to others, potentially identifying it as part of the training set based on output discrepancies or changes when the point is included or excluded from training data. The rising trend of data sharing, particularly sensitive personal information, among organizations underscores the critical need for stringent privacy measures, guided by regulations such as the General Data Protection Regulation (GDPR) [18] and the California Consumer Privacy Act (CCPA) [19]. These regulations are pivotal in upholding data privacy, especially within transfer learning where knowledge transfer occurs across varied organizational echelons [20], [21], [22], [23], [24], [25], [26].

In transfer learning, research efforts like those from Zou et al. and Hidano et al. [27], [28] have focused on black-box MIAs. Zou et al. [27] were pioneers in this field, applying shadow training to analyze MIAs against the target model, albeit without extending this analysis to the teacher model. Hidano et al. [28] advanced this research by introducing transfer shadow training, enhancing attack accuracy by exploiting the transferred model's parameters. However, neither study fully explored the critical relationship between teacher and student models in a white-box way, a key aspect for a comprehensive understanding of MIAs in transfer learning. Recognizing and exploring this relationship is crucial for a deeper and more accurate assessment of privacy risks in this domain. Thus, we are motivated by the question of how a white-box attack framework, attentively analyzing the complex interactions between teacher and student models, can yield a

Manuscript received 13 March 2024; revised 13 May 2024; accepted 4 June 2024. Date of publication 12 June 2024; date of current version 3 July 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB2700200, in part by the National Natural Science Foundation of China under Grant 62076187, in part by the Key Research and Development Program of Hubei Province under Grant 2021BAA190 and Grant 2022BAA039, and in part by the Key Research and Development Program of Shandong Province under Grant 2022CXPT055. The associate editor coordinating the review of this article and approving it for publication was Dr. Paolo Gasti. *(Corresponding author: Jing Chen.)*

Cong Wu and Yang Liu are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: cong.wu@ntu.edu.sg; yangliu@ntu.edu.sg).

Jing Chen, Qianru Fang, and Kun He are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: chenjing@whu.edu.cn; fangqr@whu.edu.cn; hekun@whu.edu.cn).

Ziming Zhao is with the Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115 USA (e-mail: z.zhao@northeastern.edu).

Hao Ren is with the School of Cyber Science and Engineering, Sichuan University, Chengdu 610017, China (e-mail: hao.ren@scu.edu.cn).

Guowen Xu is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: guowenxu@cityu.edu.hk).

Yang Xiang is with the Digital Research, Swinburne University of Technology, Hawthorn, VIC 3122, Australia (e-mail: yxiang@swin.edu.au).

Digital Object Identifier 10.1109/TIFS.2024.3413592

more comprehensive and nuanced understanding of privacy risks in transfer learning environments.

A. Our Approach

We propose a new MIA vector in transfer learning to determine if a data point was part of the teacher model's training by analyzing the student model. In our *white-box* scenario, the attacker has full access to the student model's internal architecture and feature representations. Although the attacker does not have direct access to the teacher model or its data, they use a shadow dataset similar to the student model's training data, classifying it into member and non-member groups to train a shadow student model [3], [7], [29], [30]. The attacker then queries the actual and shadow student models using the shadow dataset to collect feature representations and analyze variations.

The attacker establishes adaptive thresholds based on these feature variations to infer membership. This strategic method exploits the differences between the actual and shadow student models' feature representations, revealing patterns indicative of membership in the teacher model's training data. By using unique data transformation discrepancies in transfer learning, our approach significantly enhances the efficacy of MIAs. This highlights privacy risks in transfer learning, especially given the increasing reliance on private or semi-private datasets, emphasizing the need for stronger privacy safeguards.

Significantly different from prior MIA researches [3], [13], [27], [28], [31], [32], [33], we unveil a novel white-box MIA vector specifically tailored for transfer learning contexts, addressing a gap not extensively covered in existing literature. This attack distinctively analyzes differential representations between teacher and student models, a step beyond the traditional focus on direct feature comparisons, thus providing a nuanced understanding of transfer-induced vulnerabilities. We also innovate with a trinary decision framework, which refines the attack precision by distinguishing among various membership statuses more accurately. Our investigation into the complex interplay between teacher and student models uncovers subtle yet critical insights into their shared vulnerabilities, advocating for enhanced privacy measures in transfer learning. Through a meticulous comparative analysis using shadow models, our work underscores the need for a reevaluated defense strategy, positioning our contributions as a pivotal reference for future research in securing transfer learning mechanisms.

B. Contribution

We summarize our contribution as follows:

- We unveil a new MIA vector in transfer learning, which is the first white-box attack against transfer learning to the best of our knowledge. It employs differential representation analysis and a ternary decision framework to elucidate privacy vulnerabilities in teacher-student model interactions.
- To elucidate the interplay between teacher and student models, we present an adaptive threshold selection mechanism, enabling the extraction of representation differences between the student model and corresponding shadow model (§IV).

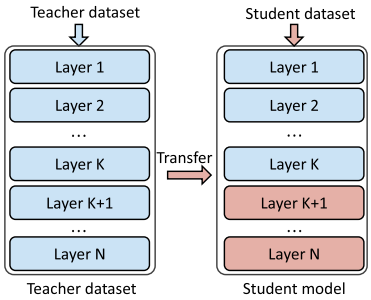


Fig. 1. Illustration of transfer learning.

- Our approach is rigorously tested across diverse settings, four datasets, and different transfer learning tasks, comparing it against current state-of-the-art methods. The results indicate that our proposed attack achieves a high attack accuracy, and more effectively exposes the teacher model's membership privacy then previous black-box MIAs (§VI).
- We systematically consider other two possible attacks under transfer learning, including inferring the student membership using the knowledge of the student model, and inferring the teacher membership using the knowledge of the teacher model. We also evaluate the performance of these attacks cases and compare them with our attack cases (§V).

II. Background

This section briefs transfer learning and MIA.

A. Transfer Learning

Transfer learning leverages a pre-trained teacher model's architectural framework and layer weights to enhance a student model's efficacy on related tasks [34], [35], [36]. Its proven effectiveness is particularly beneficial for entities with limited data or computational power, enabling the development of precise models with optimized resource utilization for task-specific applications.

As illustrated in Figure 1, transfer learning begins with the student model adopting the teacher model's feature extraction layers and appending a new dense layer to fit its specific task. In subsequent training phases, the student model, using a smaller dataset, retains fixed weights in the initial K feature extraction layers, while adapting the weights in the remaining layers. These extraction layers are crucial for recognizing input data patterns through convolution and pooling, while the dense layers synthesize this information for specific outputs or classifications. Maintaining the pre-trained layers unchanged capitalizes on their learned features, optimizing training efficiency and reducing resource demands. The extent of layer freezing, determined by the similarity between the teacher's and student's tasks, varies from minimal adjustments for closely related tasks to more extensive tuning for distinct ones, ensuring the student model's relevancy and performance.

B. Membership Inference Attack

MIA aims to ascertain if a deep learning model, often involving complex nonlinear functions and numerous layers, was trained utilizing a given data record [3], [37].

This process of inquiry requires the attacker to interact with the model, supplying the given data point and observing the output to assess the membership status of the provided data. This assessment, a key facet of ML privacy research, can be formally represented as in Eq. 1. The primary attack task is to construct a MIA model $\mathcal{A}$ that operates against target model $\mathcal{M}$, with the aim to predict the membership status of a given data point $\mathbf{x}_{target}$.

$$\mathcal{A}(\mathbf{x}_{\text{target}}, \mathcal{M}, \Omega) = \begin{cases} 1, & \text{if } \mathbf{x}_{\text{target}} \text{ is in the training data of } \mathcal{M} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The target model $\mathcal{M}$, a pre-trained and possibly heavily parametrized model, and the prior knowledge Ω, often encapsulating assumptions regarding the data distribution or the target model's characteristics, can be accessed by the attacker. The attack model $\mathcal{A}$ serves as a binary classifier, tasked with determining whether the given data point originated from the training set. It outputs the indicative of the fact that the input data record $\mathbf{x}_{target}$ is a constituent of the target model $\mathcal{M}$'s training dataset.

In terms of the level of insight an attacker has into the target model and its training data, we can distinguish between two forms of MIAs: black-box and white-box. In the scenario of a black-box attack, the attacker's access is limited to output of target model. This permits them to extract the prediction vectors, namely the softmax output, but their insight ends here. Shifting to the white-box attack, the situation is vastly different. The attacker enjoys extensive access, stretching to the complete details of the model: the weights, architecture, and at times, even the gradients. This increased accessibility endows the attacker with a more holistic comprehension of the functioning of the model. To ensure a smooth transition between these two scenarios, it's vital to highlight that while the black-box approach restricts the attacker's knowledge, the white-box approach amplifies it, encompassing even the minutest of model details. Thus, the threat level in a white-box attack escalates due to the comprehensive understanding it affords the attacker.

III. Attack Overview

In this section, we present the attack scenario, then illustrate threat model and problem formulation.

A. Attack Scenario

In practical scenarios, base models are often not trained solely on publicly available datasets like ImageNet. Instead, they might incorporate proprietary data to cater to specialized and sensitive applications like fraud detection using financial records [38], [39], disease diagnosis with personalized medical imaging [40], [41], or biometric recognition in private systems [42], [43], [44]. Commercialized models often include meta-information about their training context that attackers can exploit. Platforms like Hugging Face,[1] Azure AI Services,[2] TensorFlow Hub[3] frequently provide dataset descriptions (e.g., ImageNet, COCO), training methodologies, hyperparameters, benchmarks, and README files outlining training purposes and fine-tuning strategies. This detailed information allows attackers to replicate the data distribution and architecture in shadow models, facilitating realistic MIAs. For instance, if a model is known to have been trained on a proprietary medical imaging dataset for disease diagnosis, attackers can simulate a similar dataset and shadow model, enabling them to compare feature representations and infer patterns indicative of the original training data. This comprehensive level of detail makes shadow model creation straightforward and heightens the risks associated with MIAs.

In the context of transfer learning, an attacker aims to determine whether specific data were part of the teacher model's training by analyzing the intermediate outputs of the student model. Despite being trained on a distinct, often smaller, dataset, the student model retains information from the teacher model through the transfer learning process. Thus, having access to the student's outputs offers clues that can reveal the original training data.

The attacker systematically queries the student model, scrutinizing its intermediate outputs to detect patterns indicative of the teacher model's training data. They can construct shadow models mimicking the student model's behavior to analyze discrepancies in feature representations. Comparing shadow and student models, the attacker identifies discrepancies and infers the likelihood of data belonging to the teacher model's training set. This approach is practical where adversaries can observe or intercept model outputs, such as in shared or distributed computing environments. Gaining insights into intermediate outputs uncovers vulnerabilities in transfer learning, emphasizing the importance of securing these outputs.

B. Threat Model

Our research expands upon this by exploring the white-box attack scenario, where the attacker can access model's intermediate output. Specifically, we consider the following attack cases that are possible in transfer learning: MIA against the teacher model while accessing the student model (At.T & Ac.S), directly MIA against the accessible teacher model (At.T & Ac.T), and directly MIA against the accessible student model (At.S & Ac.S), thereby providing a comprehensive investigation into the vulnerabilities across transfer learning frameworks.

- **At.T & Ac.S.** The attacker has prior knowledge of the student model's parameters, structure, parameters, and training data distribution, and also has access to the student model itself. The goal is to decide if a particular data point was used in the training process of the teacher model, without having any knowledge of the teacher model. To achieve this, the attacker may utilize the shadow model dataset and the student model to generate feature sets and adaptive thresholds for MIAs.
- **At.T & Ac.T.** The attacker has none knowledge of the student model, but has prior knowledge of the teacher model.

[1] https://huggingface.co/models

[2] https://azure.microsoft.com/en-us/products/ai-services/

[3] https://www.tensorflow.org/hub

TABLE I
COMPARISON OF THREE DIFFERENT ATTACK CASES

Attack type	Target	Knowledge	Formulation
At.T & Ac.S	Teacher	Student	$\mathcal{A}_{At.T\&Ac.S}(\mathcal{M}_s(x_{target}), \Omega_s) \rightarrow \{0, 1, 2\}$
At.S & Ac.S	Student	Student	$\mathcal{A}_{At.S\&Ac.S}(\mathcal{M}_s(x_{target}), \Omega_s) \rightarrow \{0, 1\}$
At.T & Ac.T	Teacher	Teacher	$\mathcal{A}_{At.T\&Ac.T}(\mathcal{M}_t(x_{target}), \Omega_t) \rightarrow \{0, 1\}$

The attacker's goal is to decide if a specific data record was used to train the teacher model.

- **At.S & Ac.S.** The attacker has prior knowledge of the student model but lacks any knowledge about the teacher model. The attacker attempts to perform MIA toward the student model, i.e., deciding if a given input data record was used to train the student model.

Despite their similar attack pipeline, the complexity of At.T & Ac.T and At.S & Ac.S attack cases, differs significantly due to the varying sizes of training data used to train the teacher and student models. The teacher model is typically trained from scratch using a substantial amount of data, while the student model leverages a comparably smaller dataset for its training, achieved through fine-tuning the teacher model. As such, it is impractical to infer the student model's membership using solely the information from the teacher model, as the dataset used for the student model's training is not part of the teacher model's training process. In this paper, we aim to concentrate on these three potential attack cases, exploring the viability of MIAs in transfer learning.

We also note that our white-box assumption models an adversary with significant, yet plausible, system penetration capabilities, aligning with scenarios where internal threats or security breaches provide access to the model's internals, a consideration crucial for rigorous security evaluations. The white-box access in our threat model is an intentional design to test the upper bounds of an attacker's capability, offering a new stringent MIA security evaluation framework; this higher threat level assumption is reasonable as it represents the full exploitation of an attacker's ability to pose serious security threats, ensuring preparedness for worst-case scenarios. In real-world scenarios, attackers can also pragmatically deduce the nature of a base model's tasks by examining the student model's attributes, supported by associated metadata or domain-specific commonalities. This information enables them to craft shadow models that reflect the teacher's training context, making it a practical assumption for understanding and mitigating potential vulnerabilities in transfer learning frameworks.

C. Attack Formulation

The notions are given as Table II. The formation of the three attack cases are given as follows.

At.T & Ac.S. We formulate the problem of At.T & Ac.S as Eq. 2:

$$\mathcal{A}_{At.T\&Ac.S}(\mathcal{M}_s(\ \mathbf{x}_{target}), \Omega_s) \rightarrow \{0, 1, 2\} \tag{2}$$

where $\mathcal{A}_{At.T\&Ac.S}$ is the attack model. $\mathcal{M}_s$ is the student model. $\mathcal{M}_s(\mathbf{x}_{target})$ is the output of student model. where the attack model $A_{At.T\&Ac.S}$ is a three-class classifier. Ω_s is the auxiliary knowledge from the student model, typically encompassing the model architecture, hyperparameters, or output behaviors observed during interactions with M_s. Attackers can deduce such information through public documentation, direct model interaction, or inferential analysis based on common practices within the domain of the student model's application. 1 represents that the data point is a member of $\mathcal{M}_t$'s training dataset. 2 represents that the data point is a member of $\mathcal{M}_s$'s training dataset. 0 represents that the data point is a non-member of $\mathcal{M}_t$'s training dataset and $\mathcal{M}_s$'s training dataset.

TABLE II
NOTATIONS IN THIS PAPER

Notation	Description
At.T & Ac.S	Attack teacher model while accessing student model
At.T & Ac.T	Attack teacher model while accessing teacher model
At.S & Ac.S	Attack student model while accessing student model
$\mathbf{x}_{target}$	Target data point
$\mathbf{x}_{train}$	Training data point
$\mathcal{M}_t$	Teacher model
$\mathcal{M}_s$	Student model
$\mathbf{D}_s^{shadow}$	Shadow student dataset
$\mathbf{D}_t^{shadow}$	Shadow teacher dataset
$\mathbf{D}^n$	Random generated non-membership dataset
$\mathcal{M}'_t$	Shadow teacher model
$\mathcal{M}'_s$	Shadow student model
D_{train}	Training dataset
σ_1, σ_2, σ_3	The three selected thresholds
$\mathcal{A}$	The attack model
Ω_s	The prior knowledge from the student model
Ω_t	The prior knowledge from the teacher model
${}^m\mathbf{D}_s^{shadow}$	The member data of shadow student model
${}^n\mathbf{D}_s^{shadow}$	The non-member data of shadow student model
${}^m\mathbf{D}_t^{shadow}$	The member data of shadow teacher model
${}^n\mathbf{D}_t^{shadow}$	The non-member data of shadow teacher model

The intuition behind this attack stems from the premise that feature representations differ between a model trained on specific data and an analogous model trained independently. In transfer learning, these disparities in feature representations between the student and a shadow student model, especially when the latter is trained on disparate data, can reveal traces of the teacher model's training dataset. This approach capitalizes on the fundamental principle that learned representations in neural networks are inherently data-dependent, thereby justifying the use of feature representation differences as a mechanism for inferring membership in the teacher model's dataset.

At.T & Ac.T. We formulate At.T & Ac.T problem as Eq.3:

$$\mathcal{A}_{At.T\&Ac.T}(\mathcal{M}_t(\mathbf{x}_{target}), \Omega_t) \rightarrow \{0, 1\} \tag{3}$$

where the attack model $\mathcal{A}_{At.T\&Ac.T}$ is a binary classifier. $\mathcal{M}_t$ is the teacher model. $\mathbf{x}_{target}$ is the data used for membership inference. $\mathcal{M}_t(\mathbf{x}_{target})$ is the output of teacher model, and Ω_t is the auxiliary knowledge from the teacher model. 1 represents that the data point is a member of $\mathcal{M}_t$'s training dataset and 0 otherwise.

At.S & Ac.S. Given that the student model is trained based on the teacher model, a portion of the structure and parameters of the model are transferred to the student model. As a result, the privacy related to membership is encoded within

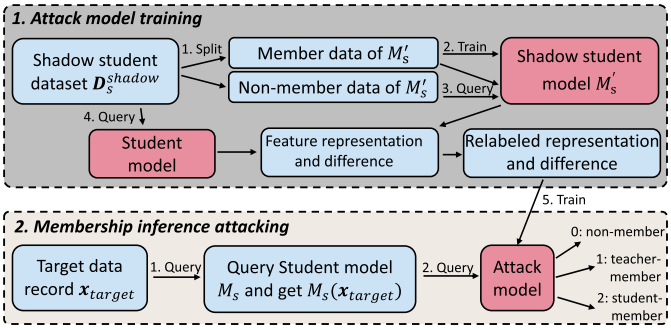


Fig. 2. Attack workflow of At.T & Ac.S.

the teacher model's parameters. Despite the retraining of the student model, the privacy implications persist. We represent the problem of At.S & Ac.S as denoted in Eq. 4.

$$\mathcal{A}_{At.S\&Ac.S}(\mathcal{M}_s(\mathbf{x}_{target}), \Omega_s) \rightarrow \{0, 1\} \tag{4}$$

where the attack model $\mathcal{A}_{At.S\&Ac.S}$ is a binary classifier. 0 represents that the data point is not a member of $\mathcal{M}_s$'s training dataset. 1 represents that the data point is a member of $\mathcal{M}_s$'s training dataset.

IV. Attack Design

In this section, we illustrate attack schemes of At.T & Ac.S.

A. Attack Model Training

As shown in Figure 2, the attacker first splits the shadow student dataset into member and non-member datasets, i.e., ${}^m\mathbf{D}_s^{shadow}$ and ${}^n\mathbf{D}_s^{shadow}$. The shadow student model $\mathcal{M}_s'$ is trained using the member dataset. Then the attacker queries the shadow student model and real student model using the member dataset respectively, and obtains the feature representation of the intermediate hidden layer, i.e., $\mathcal{M}_s'(\mathbf{x})$ and $\mathcal{M}_s(\mathbf{x})$, $\mathbf{x} \in {}^m\mathbf{D}_s^{shadow}$. Then the attacker calculates the L_2 distance between the two feature representations, i.e., $|\mathcal{M}_s'(\mathbf{x}) - \mathcal{M}_s(\mathbf{x})|$, $\mathbf{x} \in {}^m\mathbf{D}_s^{shadow}$, which represents the privacy of teacher dataset.

Typically, the tasks of the teacher and student models share substantial similarities, making the inference of the teacher model's membership challenging once the student model is trained based on the teacher model. For instance, the teacher model might classify images of cats and dogs, while the student model could be tasked with classifying different breeds of cats. In such scenarios, it is considerably tough for an attacker to infer the teacher model's membership when querying the model with non-membership data.

The attacker also generates random noisy images, $\mathbf{D}^n$ where the generated images are without semantics. The attacker queries the student model using the generated images and gets the feature representation of the intermediate layer $\mathcal{M}_s(\mathbf{x}), \mathbf{x} \in \mathbf{D}^n$. The feature representation of the intermediate hidden layer is then labeled with different labels. Specifically, the output feature representation queried using the member dataset of the shadow student model is labeled with 2, i.e., $(\mathcal{M}_s(\mathbf{x}), 2), \mathbf{x} \in {}^m\mathbf{D}_s^{shadow}$ is labeled with 2. The difference is calculated as the privacy of the teacher model member data, and is labeled with 1, i.e., $(|\mathcal{M}_s'(\mathbf{x}) - \mathcal{M}_s(\mathbf{x})|, 1), \mathbf{x} \in {}^m\mathbf{D}_s^{shadow}$. Similarly, the student model is queried using the randomly generated images, and the output feature representation of the intermediate layer is labeled with 0, i.e., $(\mathcal{M}_s(\mathbf{x}), 0), \mathbf{x} \in \mathbf{D}^n$. The attacker then uses the labeled data to train the membership inference model. The model is trained to distinguish three different classes, i.e., teacher-member, student-member, and non-member.

Algorithm 1 Training Attack Model

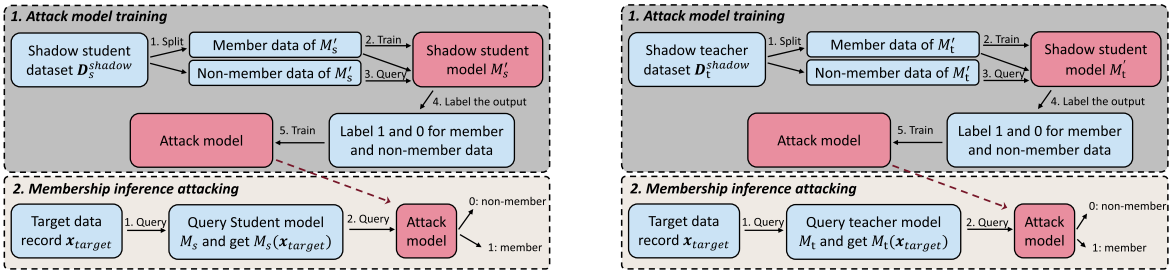
Input: $\mathbf{D}_s^{shadow}$, $\mathbf{D}^n$, $\mathcal{M}_s$
Output: The attack model $\mathcal{A}$
Split $\mathbf{D}_s^{shadow}$ into ${}^m\mathbf{D}_s^{shadow}$ and ${}^n\mathbf{D}_s^{shadow}$
Train the shadow model $\mathcal{M}_s'$ using ${}^m\mathbf{D}_s^{shadow}$
Compile the teacher-member dataset $(|\mathcal{M}_s(\mathbf{x}) - \mathcal{M}_s'(\mathbf{x})|, 1)$, $\mathbf{x} \in {}^m\mathbf{D}_s^{shadow}$
Compile the student-member dataset $(\mathcal{M}_s(\mathbf{x}), 2)$, $\mathbf{x} \in {}^m\mathbf{D}_s^{shadow}$
Compile the non-member dataset $(\mathcal{M}_s(\mathbf{x}), 0)$, $\mathbf{x} \in \mathbf{D}^n$
Train the attack model $\mathcal{A}$
return $\mathcal{A}$

Algorithm 1 details the attack model training process. The shadow student dataset $\mathbf{D}_s^{shadow}$ is divided into the member dataset ${}^m\mathbf{D}_s^{shadow}$ and the non-member dataset ${}^n\mathbf{D}_s^{shadow}$ using a 1:1 ratio; specifically, 50% of $\mathbf{D}_s^{shadow}$ is randomly selected to train the shadow model, forming ${}^m\mathbf{D}_s^{shadow}$, while the remaining 50% constitutes ${}^nD_s^{shadow}$, representing data that the shadow model has not encountered. The threshold σ_1 critically impacts the classification accuracy by dictating the distinction between member and non-member data. An optimal σ_1 minimizes both false positives and negatives, thereby enhancing the precision and recall of the attack model $\mathcal{A}$. As an example, Figure 4 present the feature representation of the student model, the shadow student model, and the feature representation differences. The feature representations and the difference is then used to decide the membership of the teacher and student models.

B. Attack Strategy

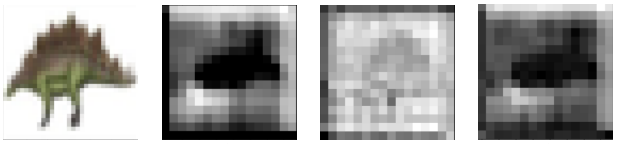
During attack phase, the attacker queries the student model using the target data record $\mathbf{x}_{target}$, and derives the feature representation of the intermediate hidden layer. Then, the attacker queries the attack model using the output feature representation and calculates the distance between the feature representation and the labeled data in the training set. L_2 distance is used as the distance metric. Specifically, there are three cases:

- (i) $dis(\mathcal{M}_s(\mathbf{x}_{target}), \mathcal{M}_s(\mathbf{x})) < \sigma_1$, $\mathbf{x} \in \mathbf{D}_s^{shadow}$. This indicates the feature representation of the target data record is similar to from the feature representation of the student shadow model member data. The target can be regarded as a student-member data point and the attack result is 2.
- (ii) $\sigma_1 < dis(\mathcal{M}_s(\mathbf{x}_{target}), \mathcal{M}_s(\mathbf{x})) < \sigma_2, \mathbf{x} \in \mathbf{D}_s^{shadow}$. This indicates that the feature representation of the target data record is not the number of student number. But it



(a) At.T & Ac.T (b) At.S & Ac.S

Fig. 3. Attack workflow of At.T & Ac.T (a) and At.S & Ac.S (b).



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

Fig. 4. A example of the raw image (a), the feature representation of the student model (b), shadow student model (c), and feature representation differences (d).

requires to further decide whether the target data record is a teacher-member data point. The attack model calculates the L_2 distance between the feature representation and the labeled teacher-member data in the training set. There are two special cases. If $dis(\mathcal{M}_s(\mathbf{x}_{target}), |\mathcal{M}_s(\mathbf{x}) - \mathcal{M}_s'(\mathbf{x})|), \mathbf{x} \in \mathbf{D}_s^{shadow} < \sigma_3$. The target data record can be regarded as a teacher-member data point. The inference attack results is 1. If $dis(\mathcal{M}_s(\mathbf{x}_{target}), |\mathcal{M}_s(\mathbf{x}) - \mathcal{M}_s'(\mathbf{x})|), \mathbf{x} \in \mathbf{D}_s^{shadow} > \sigma_3$. The target data record can be regarded as a non-member data point. The inference attack results is 0.

- (iii) $dis(\mathcal{M}_s(\mathbf{x}_{target}), \mathcal{M}_s(\mathbf{x})) > \sigma_2, \mathbf{x} \in \mathbf{D}_s^{shadow}$. This indicates the feature representation of the target data record is different from the feature representation of the shadow model member data. The inference attack results is 1.

In our approach, we use a distance metric $dis(\mathcal{M}_s(\mathbf{x}_{target}), \mathcal{M}_s(\mathbf{x}))$ to measure how closely a target sample's output from the student model aligns with known samples. This measurement is crucial for determining whether a sample is likely part of the model's training set. We set thresholds to categorize these distances, aiding in accurately distinguishing between members and non-members. These thresholds are carefully determined to ensure our attack can effectively identify membership status, demonstrating the practicality and ingenuity of our method within transfer learning's unique framework.

C. Threshold Selection

In the At.T & Ac.S scenario for conducting MIAs, we establish three distinct thresholds to accurately determine if a data record was part of the training set. Our threshold selection algorithm outlines this process meticulously. For σ_1, we process the student shadow dataset through the student model, mirroring the student model's data distribution. We then calculate the distance between feature representations from the shadow student model and the actual student model data.

Algorithm 2 Threshold Selection

Input: ${}^m\mathbf{D}_s^{shadow}$, $\mathbf{D}^n$, $\mathcal{M}_s$, and $\mathcal{M}_s'$
Output: The selected thresholds: σ_1, σ_2, σ_3

```
for x in ^mD_s^shadow do
    distance_1 ⟵ dis(M_s'(x), M_s(x))
    σ_1 ⟵ median(distance_1)
end
for x_1, x_2 in D^n, ^mD_s^shadow do
    distance_2 ⟵ dis(d(M_s(x_1), M_s(x_2)))
    σ_2 ⟵ median(distance_2)
end
for x_1, x_2 in D^n, ^mD_s^shadow do
    distance_3 ⟵ dis(M_s(x_1), |M_s(x_2) − M_s'(x_2)|)
    σ_3 ⟵ median(distance_3)
end
return σ_1, σ_2, σ_3
```

The median value of these distances is chosen as the threshold σ_1, effectively classifying membership status. The use of median values for threshold selection is based on their robustness against outliers and their accurate representation of central tendencies in data distributions. Medians provide a stable threshold, less influenced by extreme values, ensuring consistent performance across diverse datasets.

For establishing the threshold σ_2, the attacker synthesizes random noisy images to query the student model, subsequently measuring the distances between these queried results and the feature representations from the shadow model's member data. The median of these calculated distances is selected as σ_2. To set σ_3, the attacker compares the feature representations obtained from the noisy image queries against the differences found between feature representations from the student model and shadow model member data. The median value from these comparative distances is then assigned as the threshold σ_3, aiding in the nuanced differentiation of membership inferences.

V. Other Two Typical Attacks

We also present two typical MIAs in transfer learning for comparing, where MIAs are considered in a white-box way.

A. Attack Scheme of At.T & Ac.T

As shown in Figure 3(a), the attacker queries the shadow teacher model using the data that participated in the shadow

model training. The representation of the intermediate layer, $\mathcal{M}'_t(\mathbf{x})$, for $\mathbf{x} \in {}^{m}\mathbf{D}_t^{shadow}$ is labeled with 1, indicating membership. Similarly, the shadow teacher model is also queried using data that did not appear in the shadow model training. The output of the intermediate layer, $\mathcal{M}'_t(\mathbf{x})$, for $\mathbf{x} \notin \mathbf{D}_{shadow}$ is labeled with 0, indicating non-membership. The attacker then uses the output of the intermediate layer and the corresponding labels to train the attack model. The goal of the attack model is to distinguish the membership status of the shadow teacher model. In the attack stage, the attacker queries the shadow teacher model using data that participate in the shadow model training. The output of the intermediate layer, $\mathcal{M}'_t(\mathbf{x})$, for $\mathbf{x} \in {}^{m}\mathbf{D}_t^{shadow}$ is labeled with 1, indicating membership. Similarly, the shadow teacher model is also queried using data that did not appear in the shadow model training. The output of the intermediate layer, $\mathcal{M}'_t(\mathbf{x})$, for $\mathbf{x} \notin {}^{m}\mathbf{D}_t^{shadow}$ is labeled with 0, indicating non-membership. The attacker then uses the output of the intermediate layer and the corresponding labels, i.e., $(\mathcal{M}'_t(\mathbf{x}), 1)$ and $(\mathcal{M}'_t(\mathbf{x}), 0)$, to train the attack model. The goal of the attack model is to distinguish the membership status of the shadow teacher model.

B. Attack Scheme of At.S & Ac.S

As shown in Figure 3(b), the attacker queries the shadow model using the data that participated in shadow model training. The output of the intermediate layer, $\mathcal{M}'_s(\mathbf{x})$, $\mathbf{x} \in {}^{m}\mathbf{D}_s^{shadow}$, is labeled with 1, i.e., belonging to the member. Similarly, the shadow model is also queried using the data not appearing in shadow model training. The output of the intermediate layer, $\mathcal{M}'_s(\mathbf{x}), \mathbf{x} \notin {}^{m}\mathbf{D}_s^{shadow}$, is labeled with 0, i.e., not belonging to the member. Then the attacker uses the output of the intermediate layer and the corresponding labels, i.e., $(\mathcal{M}'_s(\mathbf{x}), 1), \mathbf{x} \in {}^{m}\mathbf{D}_s^{shadow}$ and $(\mathcal{M}'_s(\mathbf{x}), 0)$, $\mathbf{x} \notin {}^{m}\mathbf{D}_s^{shadow}$, to train the binary classification model. The model is trained to distinguish the members of the data from the outputs of the intermediate layer. Similarly, in the attack stage, the adversary queries the student model using the target data record and get the feature representation of hidden layer. Then, the attacker queries the attack model using the feature representation and obtains the predicted membership status.

VI. Experimental Results

In this section, we report performance results.

A. Evaluation Setup

1) Datasets: We used the following widely-used datasets in previous MIA works for evaluation [13], [27], [28], [31], [45]:

- *ImageNet* [46] serves as a foundational dataset in computer vision, featuring extensive classes and images, commonly used to train teacher models in transfer learning. The pre-trained models are then adapted to student models for tasks in similar domains, leveraging its detailed class structure.

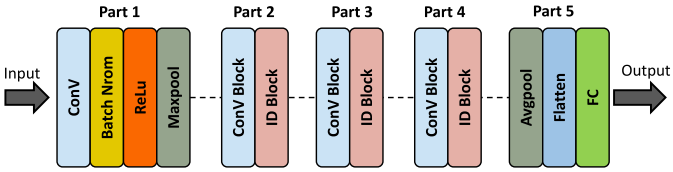


Fig. 5. The structure of ResNet50, we divided the model into five parts for transfer learning in evaluation.

- *CIFAR-100* [47] comprises 60,000 three-channel color images across 100 varied classes like flowers and fish, each with 500 training and 100 testing images of 32×32 pixels.
- *Flowers102* [48] features 102 flower types, totaling 7,169 images distributed across 102 classes with an imbalanced dataset ranging from 40 to 258 images per class. The dataset includes 6,149 training and 1,020 testing images.
- *Cats vs. Dogs* [49] dataset includes 25,000 training and 12,500 testing images, evenly divided between cat and dog.

2) Experimental Settings: ImageNet was used as the teacher dataset to pre-train as teacher model as it contains various classes and has the largest number of images. During our evaluation, we trained the teacher model using ResNet50, VGG19, Inception v3, and DenseNet169 respectively. We reimplemented the CNN architectures used in our experiments under TensorFlow 1.15.2.

In Figure 5, we illustrate the widely-used ResNet50 model to evaluate performance of MIAs. The model is segmented into five key parts: Part 1 includes initial convolution, normalization, ReLU activation, and max pooling layers; Parts 2 to 4 contain progressively complex Identity (ID) Blocks for feature extraction; and Part 5 concludes with average pooling and a fully connected (FC) layer for output generation. In our experiments, Parts 1-3 of the teacher model were immobilized to initialize the student models. We fine-tuned these student models using the CIFAR-100, Flowers102, and Cats vs. Dogs datasets, each inheriting pre-trained weights from the teacher model. Additionally, we evaluated various model architectures, including VGG19, InceptionV3, and DenseNet169, experimenting with different freezing configurations (e.g., Parts 1-2, 1-3, and 1-4) to observe the impact on transfer learning effectiveness and subsequent inference attack accuracy.

To ensure the evaluation fairness, we evenly divided the teacher/student datasets into member and non-member groups, ensuring no overlap and a balanced 1:1 ratio for unbiased analysis. Given that the teacher and student datasets originate from distinct sources, their different distributions enhance the assessment's robustness. For constructing shadow models, we split their corresponding datasets into 70% for training and 30% for testing, tailoring unique models to each student task. Evaluations were conducted by querying the models with a balanced mix of 10% member data and an equal amount of non-member data from the testing sets. This process was repeated across ten iterations for each attack scenario to derive consistent performance metrics.

3) Metrics: MIA can be perceived as a binary classification challenge: deciding whether a given data point belongs to the target model's training set (positive) or not (negative).

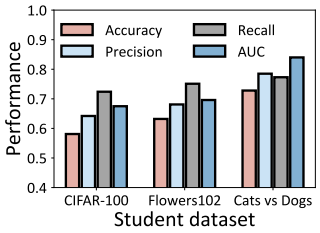


Fig. 6. Performance of At.T & Ac.S under different datasets.

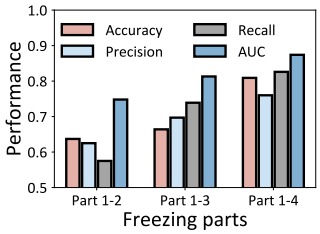


Fig. 7. Performance under varied freezing parts.

We used the widely-used metrics, including precision, recall, accuracy, and the area under receiver operating characteristic curve (AUC). Recall is the ratio of accurately classified member data to total member data. Precision is the ratio of correctly classified member data to all predicted member data. Accuracy is the ratio of correctly classified data points to all data points. AUC measures the likelihood that the prediction score of member data exceed non-member data.

Our evaluations were conducted on Tesla P4*3 GPUs using TensorFlow 2.5.0, with each model trained for 100 epochs, a learning rate of 0.001, and a batch size of 32, utilizing Adam as the optimizer. We partitioned the initial dataset, allocating 70% for training the teacher model and 30% for the student model, predominantly using the ResNet50 architecture.

B. Overall Performance of At.T & Ac.S

We evaluated the performance of the proposed At.T & Ac.S of using the input feature representation to perform MIA. The teacher model was trained using ImageNet dataset and ResNet50 as the base teacher model. We transferred the teacher model to the student model using CIFAR-100, Flowers102, and Cats vs Dogs dataset, respectively. Specifically, we froze part 1-3 (Figure 5) to train the student model when performing transfer learning. To perform the attack, the shadow student model was trained from scratch using the shadow student dataset, where the hidden layer did not contain the behavioral characters of the teacher model.

Figure 6 presents the evaluation results of At.T & Ac.S on three student datasets: CIFAR-100, Flowers102, and Cats vs Dogs. Our method achieves accuracies of 0.581, 0.632, and 0.728, respectively, with AUC values approaching or exceeding 0.7, indicating effectiveness beyond random guessing. The proposed MIA remains effective across most teacher model classes in transfer learning. Notably, the Cats vs Dogs dataset demonstrates higher attack accuracy (0.728) and precision (0.785) compared to CIFAR-100 (0.581, 0.642) and Flowers102 (0.632, 0.681). We speculate this is due to Cats vs Dogs having fewer classes and data points, resulting in smaller changes during transfer learning.

C. Impact of the Number of Frozen Layers

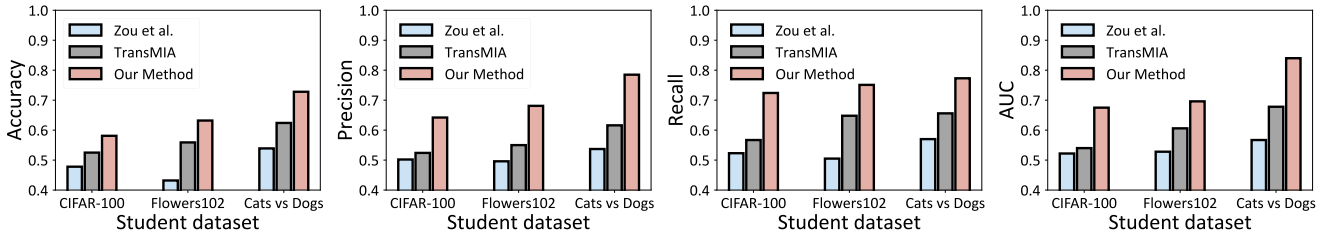
To evaluate attack performance, we varied the number of frozen layers in the student model, which was trained by freezing different parts of the ResNet50 teacher model, initially trained on the ImageNet dataset. Specifically, ResNet50 was divided into five parts (Figure 5), and the student model was trained by freezing parts 1-2, 1-3, and 1-4 respectively. We transferred the model trained on the ImageNet dataset to classify the Cats vs Dogs dataset. Figure 7 shows that MIA accuracy improves with more layers frozen: 0.637 for parts 1-2, 0.664 for parts 1-3, and 0.809 for parts 1-4. We speculate that freezing more layers preserves more abstract features, allowing the attacker to better distinguish data records from the teacher model's dataset. Freezing more layers during transfer learning typically preserves pre-learned features, potentially enhancing task accuracy for similar tasks, but may limit adaptability for tasks with different nuances.

Additionally, we also experimented with the CIFAR-100 and Flowers102 datasets. For the CIFAR-100 dataset, we observed a consistent trend with the Cats vs Dogs dataset, where attack performance improved as more layers were frozen, with accuracy varying from 0.527 to 0.681 across different frozen parts. In the Flowers102 dataset, the results were less pronounced but still indicated better attack performance with more frozen layers, with accuracy varying from 0.582 to 0.652. These observations suggest that the relationship between the number of frozen layers and attack efficacy is generally applicable, though the intensity of the effect may vary depending on the dataset characteristics.

D. Comparison With SOTA Methods

We also conducted a systematic comparison of our proposed MIA with SOTA methods, including Zou et al. [27] and TransMIA [28]. Zou et al. [27] conducted an empirical study of MIAs against transfer learning using shadow training techniques, but did not consider the interconnectedness between the teacher and student models. TransMIA [28] proposed a transfer shadow training method for implementing MIAs against transfer learning, where the attacker creates the shadow model by leveraging the parameters of the transferred model. Specifically, we compared the performance of using intermediate feature representations to perform MIAs. Typically, we froze first three layers (Figure 5) when performing transfer learning to the student model, and the shadow student model was trained from scratch using the shadow student dataset.

As shown in Figure 8, the performance of our proposed approach is compared to SOTA methods on three datasets. The results indicate that our method outperforms existing methods on all three datasets. For instance, on the student classification task of the cats vs dogs dataset, our method achieves an accuracy of 0.728, which is superior to the accuracy of 0.624 of TransMIA and 0.539 of Zou et al. On the student classification task of the CIFAR-100 dataset,



(a) Accuracy (b) Precision (c) Recall (d) AUC

Fig. 8. Comparison of SOTA methods and our method on MIA accuracy (a), precision (b), recall (c), and AUC (d).

TABLE III
PERFORMANCE UNDER DIFFERENT TEACHER-STUDENT MODEL

Base model	Accuracy	Precision	Recall	AUC
ResNet50 - CIFAR-100	0.581	0.642	0.724	0.675
ResNet50 - Flowers102	0.632	0.681	0.751	0.696
ResNet50 - Cats vs Dogs	0.728	0.785	0.773	0.840
VGG19 - CIFAR-100	0.545	0.543	0.572	0.580
VGG19- Flowers102	0.587	0.591	0.566	0.631
VGG19 -Cats vs Dogs	0.627	0.632	0.610	0.684
Inception v3 - CIFAR-100	0.568	0.557	0.669	0.608
Inception v3 - Flowers102	0.633	0.621	0.681	0.704
Inception v3 - Cats vs Dogs	0.700	0.707	0.684	0.774
DenseNet169 - CIFAR-100	0.593	0.584	0.650	0.659
DenseNet169 - Flowers102	0.671	0.671	0.669	0.739
DenseNet169 - Cats vs Dogs	0.725	0.731	0.711	0.801

our method achieves an accuracy of 0.581, which surpasses the accuracy of 0.525 of TransMIA and 0.478 of Zou et al. Overall, our proposed MIA is able to expose more membership privacy of the teacher model than SOTA methods.

E. Impact of Different Teacher Models

We evaluated the impact of different teacher models on the performance of our proposed MIA using four different teacher models. Specifically, we used VGG19 [50], ResNet50 [51], Inception v3 [52], and DenseNet169 [53] as the teacher model structure respectively, since these models achieve the impressive performance and widely used for teacher model training [54]. These models were trained on the ImageNet dataset and transferred to three student datasets. Note that, we used the same student model structure as the teacher model structure. When transferring the teacher model to the student model, we froze the first half of the model so as not to participate in the model training, while the back half of the model is included in the model updating. Specifically, we froze the first three blocks for VGG1, part 1-3 for ResNet50, module `3×Inception` and `5×Inception` for Inception v3,and the first two dense blocks for DenseNet169.

Table III presents the accuracy, precision, recall, and AUC under four teacher models, i.e., VGG16, ResNet50, Inception-v3, and DenseNet169, and three datasets, i.e., CIFAR-100, Flowers102, and Cats vs Dogs. The results indicate that our proposed method can be generalized to different teacher models. For instance, when transferring to the CIFAR-100 dataset, ResNet50, VGG19, Inception v3, and DenseNet169 achieve an accuracy of 0.581, 0.545, 0.568, and 0.593, respectively. Similarly, when transferring to the Cats vs Dogs dataset, ResNet50, VGG19, Inception v3, and DenseNet169 achieve an accuracy of 0.728, 0.627, 0.700, and 0.725, respectively. Additionally, it can be observed that VGG is less susceptible to membership inference compared to other model structures in transfer learning. One possible explanation for this phenomenon may be that VGG is a model with more parameters than the other models, and the parameters of VGG are more challenging to train using the same setting.

TABLE IV
ATTACK PERFORMANCE OF AT.T & AC.T UNDER FOUR DIFFERENT MODEL STRUCTURES

Base model	Accuracy	Precision	Recall	AUC
ResNet50	0.732	0.773	0.864	0.869
VGG19	0.652	0.649	0.665	0.714
Inception v3	0.751	0.766	0.725	0.830
DenseNet169	0.769	0.794	0.729	0.849

F. Performance of At.T & Ac.T and At.S & Ac.S

1) Evaluation of At.T & Ac.T: As mentioned before, in this attack, the teacher model can be accessed by the adversary, and the goal is to decide whether the input data record is used to train the student model, which is similar to a typical MIA. We evaluated the performance of At.T & Ac.T using ImageNet as the teacher dataset under four typical models, including VGG19, ResNet50, Inception v3, and DenseNet169.

Table IV shows accuracy, precision, recall, and AUC of At.T & Ac.T under four teacher models, i.e., VGG16, ResNet50, Inception-v3 and DenseNet169, The results show that our method achieves an attack accuracy of 0.732, 0.652, 0.751, and 0.769 under the four teacher models, respectively. The attack AUC is higher than 0.7, e.g., 0.869, 0.714, 0.830, and 0.849. Besides, it can be observed that the attack accuracy on ResNet50, Inception v3, and DenseNet169 is higher than VGG19, where VGG19 is only with a MIA accuracy of 0.652.

In addition, we speculate that attack accuracies can be significantly affected by the dataset overlap; for example, a teacher model trained on the broad ImageNet and transferred to the specialized Flowers102 might exhibit higher attack accuracies due to the pronounced focus and enhanced feature learning on flower classes, underscoring the role of data relevance and specificity in the transfer learning context for membership inference success.

2) Evaluation of At.S & Ac.S: We evaluated the performance of At.S & Ac.S using three student datasets, including CIFAR-100, Flowers102, and Cats vs Dogs, under the teacher

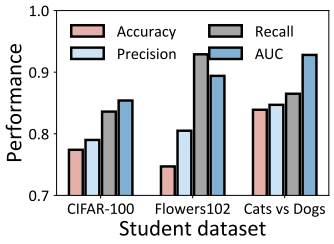


Fig. 9. Attack performance of At.S & Ac.S under CIFAR-100, Flowers102, and Cats vs Dogs dataset.

model of ResNet50 trained using ImageNet. We transferred the teacher model to the three different tasks, respectively. Specifically, part 1-3 (Figure 5) of the teacher model was frozen to train the student model on the student dataset. Figure 9 shows accuracy, precision, recall, and AUC of At.S & Ac.S under different student datasets. The results show that our method achieves a high attack accuracy of 0.774, 0.747, and 0.839 on CIFAR-100, Flowers102, and Cats vs Dogs, respectively. The overall attack AUC is higher than 0.85, e.g., 0.854, 0.894 and 0.928 on three datasets. We also observed that the attack performance on Cats vs Dogs dataset is higher than CIFAR-100, Flowers102. We speculated that the reason is that the classification task on Cats vs Dogs is more simple than the other two datasets, and thus has a higher classification accuracy when transferring the teacher model to the student task.

VII. Related Work

A. Memebership Inference Attack

MIAs have been the subject of a significant amount of research, and it has been shown that they can compromise the privacy of the training dataset of the target model [15], [37], [55], [56], [57]. Existing MIAs can be roughly divided into two categories: classifier-based and metric-based methods.

1) Classifier-Based Methods: Classifier-based methods in MIAs use binary classifiers trained through shadow training to mimic the target model's behavior and predict data point membership status, effectively utilizing training and testing data to generate informative datasets for inference [3], [7], [58]. Shokri et al. [3] were pioneers, employing shadow training for MIAs with black-box access, setting a foundational methodology for others like Salem et al. [7], who optimized this approach to enhance attack accuracy. Meanwhile, Long et al. [58] introduced strategic targeting within MIAs, underscoring the strategic nuance in identifying and exploiting model vulnerabilities.

2) Metric-Based Methods: Metric-based methods in MIAs rely on analyzing prediction vector metrics against specific thresholds to ascertain membership status [15], [57]. Yeom et al. highlighted a novel MIA approach focusing on prediction accuracy and loss, indicating that risks extend beyond model overfitting [15]. Song et al. critiqued existing entropy-based MIAs for neglecting true label characteristics, proposing a refined method integrating prediction entropy with the actual label for more accurate inference [57]. Similarly, Hui et al. introduced a shadow-less referenceless membership inference utilizing prediction entropy to gauge data's training involvement without shadow models [59]. Li et al. embraced a black-box framework, leveraging adversarial perturbation to maintain efficacy even with limited model output [60], while Liu et al. adopted a distilled loss trajectory approach for membership detection, analyzing losses across model training stages [61].

Recent research has examined white-box MIA scenarios, where attackers obtain comprehensive access to the target machine learning model's architecture and parameters. Nasr et al. [62] discovered that while final predictions and intermediate computations alone did not surpass black-box attack accuracies, integrating gradients of the prediction loss relative to model parameters significantly improved attack effectiveness in white-box contexts. Similarly, Jayaraman et al. [63] proposed an MIA that leveraged the loss variability caused by perturbations, showing its effectiveness in datasets with class imbalances. Besides, MIAs have been documented in adjacent domains like P2P federated learning and clustered federated learning, where transfer learning features are utilized. For instance, Luqman et al. [64] demonstrated that federated learning structures are susceptible to MIAs due to the shared learning frameworks.

B. MIAs Against Transfer Learning

Transfer learning, crucial for both industrial and academic progress [35], [36], faces significant security risks, including data poisoning [65], backdoor attacks [66], and susceptibility to adversarial examples [1]. Within this context, the exploration of MIAs by Zou et al. [27] and Hidano et al. [28] has advanced understanding of these vulnerabilities, though limitations in attack accuracy remain due to insufficient exploitation of complex teacher-student model interrelations.

Zou et al. [27] conducted an empirical investigation into MIAs under transfer learning, facing challenges in achieving high accuracy in black-box attacks, as shown in their CIFAR-100 dataset results. Hidano et al. [28] improved attack effectiveness using a white-box approach with a transfer shadow training strategy, but this required specific shadow model configurations of the teacher dataset. Our research builds on these studies, adopting a nuanced white-box framework that thoroughly examines differential feature representations between teacher and student models. By dissecting the intricacies of knowledge transmission and representation, our approach aims to enhance the precision of membership inference, expanding the discussion on securing transfer learning frameworks against such attacks.

Our approach leverages the interplay between teacher and student models, analyzing hidden layer representation discrepancies to infer membership status. Unlike conventional black-box methodologies that may overlook this interconnection, our white-box-centric method, adaptable to black-box scenarios via surrogate modeling, delves deeper into model internals for more insightful analysis. Prior studies, such as Zou et al. [27], may not fully address teacher-student model interconnectivity due to black-box constraints. In contrast, our method enhances attack efficacy in real-world transfer learning contexts by capitalizing on this aspect. Additionally,

TABLE V
COMPARISON OF OUR APPROACH AND EXISTING WORKS

Aspect	Our Approach	Zou et al.[27]	TransMia[28]
Scenario	black-box	Black-box	Black-box
Focus	Teacher-student model	General privacy attacks	Student model
Methodology	Hidden layer representation analysis	Shadow model training	Only student model analysis

we advance beyond TransMia [28] by integrating insights from both teacher and student models, rather than isolating the student model analysis. Table V presents a high-level comparison of our approach and existing works.

VIII. DISCUSSION

This section discusses mitigation strategies and limitations.

A. Attack Mitigation Strategies

As of now, specific defense mechanisms to mitigate MIAs in transfer learning are not well-defined [45], limiting our ability to compare our study against established MIA defenses and highlighting an urgent research gap. Future investigations are imperative to devise and assess tailored defenses that can effectively protect privacy within transfer learning contexts.

In this light, we outline potential defense strategies to address MIAs in transfer learning environments: *(i) Output randomization and differential privacy*, applying noise addition to model parameters and data or using subset-based training and prediction thresholding [67], while it may reduce the model performance; *(ii) Adversarial training*, integrating adversarial examples into training to fortify defenses [33]; *(iii) Generative learning approaches*, leveraging generative adversarial networks to obscure membership signals; *(iv) Model splitting*, distributing training across various models on disjoint data subsets to obfuscate membership clues; and *(v) Model pruning*, trimming unnecessary parameters to minimize information leakage and potentially pruning the teacher model pre-transfer to attenuate the linkage with the student model, thereby enhancing privacy protections.

B. Limitation and Future Work

Despite our diligent efforts to maintain the validity of our study, certain limitations remain. We focused solely on image classification tasks, omitting other domains such as natural language processing, medical analysis, and object detection. Additionally, our attack method targets typical parameter-based transfer learning, excluding other types such as instance-based, feature representation-based, and relational knowledge-based transfer learning [68], [69], [70], [71]. Evaluating membership inference attacks across various transfer learning methods and tasks is crucial. Our study mainly considers attacks on the teacher model with prior knowledge from the student model. Future research should explore attack performance when examining prediction vectors for different classes. While we used Euclidean distance as the primary metric, other metrics like Manhattan and Cosine distances showed lower MIA accuracy; thus, investigating alternative metrics, such as Mahalanobis distance, is essential. We observed that freezing larger parts of the model increases recall and performance, which requires further mathematical substantiation in future work.

1) Assumptions: While our model assumes white-box access, offering a stringent evaluation context, the comparison with existing works [27], [45], [72]'s black-box scenario is to illustrate how model accessibility influences attack success. Acknowledging the potential difference in real-world shadow and teacher/student dataset distributions, our experiment uses the same dataset to underscore our approach's effectiveness under less-than-ideal conditions. Future research will assess this in more varied scenarios.

Our attacker model, grounded in white-box assumptions, reflects practical adversarial capabilities in contemporary cybersecurity landscapes, where attackers can access or infer detailed model information. This model's practicality is substantiated by real-world scenarios wherein entities might gain insider information or exploit system vulnerabilities to access model details, elevating our approach beyond traditional black-box methods. By delving into the nuanced interconnectivity between teacher and student models, our methodology not only showcases superior performance over existing black-box approaches like [27], [28] but also demonstrates a pragmatic understanding of attack vectors, thereby enhancing the relevance and applicability of our findings in addressing and mitigating real-world transfer learning vulnerabilities.

In addition, while attack mitigation strategies can reduce attack efficacy, it is important to consider their impact on model performance. Future research should focus on balancing privacy protection with accurate predictions, especially in sensitive applications where membership inference attacks pose significant risks. Also, future work will involve a detailed empirical investigation into the dimensionality and feature characteristics of different datasets to better understand their impact on attack performance.

To adapt our methodology to the black-box setting, attackers can use iterative queries to approximate the target model's behavior through a shadow model, which mimics the target model's responses. By analyzing discrepancies in output probabilities and refining the shadow model, attackers can identify membership patterns via probabilistic analysis. This adaptation necessitates careful threshold tuning to maximize inference accuracy without direct access to internal model parameters. In a typical teacher $\rightarrow$ student $\rightarrow$ student transfer scenario, each model retains features from the previous one, preserving patterns indicative of the original teacher's training data. Our methodology can effectively trace these patterns, even across multiple transfer levels. Future research should investigate the resilience of this attack approach in complex transfer chains to strengthen privacy defenses.

2) Model Structure: Our evaluation utilizes popular transfer learning models, including ResNet50, VGG19, Inception v3, and DenseNet169 [36], [73], selected for their optimal balance of computational efficiency and robust performance in real-world applications. Their depth and complexity enable nuanced analysis of hidden layer representations between teacher and student models. By leveraging deep learning prin-

ciples of feature representation and transferability, the method is robust and generalizable across architectures with hierarchical feature processing. Preliminary tests using ResNet50 have yielded promising results, underscoring the method's potential. To confirm its broad applicability and adaptability, future research should evaluate additional models, e.g., Inception v2 and EfficientNet, and different frameworks, e.g., transformer across varying architectures and complexities.

IX. Conclusion

In this study, we introduce a new MIA methodology in transfer learning, elucidating the teacher model's privacy vulnerabilities. By analyzing the nuanced interplay between teacher and student models, our approach effectively identifies and leverages differences in feature representations to infer the knowledge transferred. Our extensive evaluations confirm the efficacy of this method, revealing its capability to uncover more significant privacy details from the teacher model compared to existing SOTA approaches. These findings underscore the critical need to account for the dynamic relationship between teacher and student models when assessing and fortifying against membership inference threats in transfer learning scenarios.

References

[1] B. Wang, Y. Yao, B. Viswanath, H. Zheng, and B. Y. Zhao, "With great training comes great vulnerability: Practical attacks against transfer learning," in *Proc. USENIX Secur.*, 2018, pp. 1–18.

[2] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, "Transfer learning in deep reinforcement learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 1–20, 2023.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE SP*, 2017, pp. 1–16.

[4] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 343–362.

[5] D. Wu et al., "Understanding and defending against white-box membership inference attack in deep learning," *Knowledge-Based Syst.*, vol. 259, Jan. 2023, Art. no. 110014.

[6] G. Liu, T. Xu, R. Zhang, Z. Wang, C. Wang, and L. Liu, "Gradient-leaks: Enabling black-box membership inference attacks against machine learning models," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 427–440, 2024.

[7] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2019.

[8] M. Zhang, N. Yu, R. Wen, M. Backes, and Y. Zhang, "Generated distributions are all you need for membership inference attacks against generative models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 4839–4849.

[9] H. Sun, T. Zhu, J. Li, S. Ji, and W. Zhou, "Attribute-based membership inference attacks and defenses on GANs," *IEEE Trans. Dependable Secure Comput.*, early access, pp. 1–18, 2024.

[10] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *Proc. ICML*, 2021, pp. 1964–1974.

[11] Z. Li and Y. Zhang, "Label-leaks: Membership inference attack with label," 2020, *arXiv:2007.15528*.

[12] Y. Long et al., "Understanding membership inferences on well-generalized learning models," 2018, *arXiv:1802.04889*.

[13] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 634–646.

[14] W. Yuan, C. Yang, Q. V. H. Nguyen, L. Cui, T. He, and H. Yin, "Interaction-level membership inference attack against federated recommender systems," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 1053–1062.

[15] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE 31st Comput. Secur. Found. Symp. (CSF)*, Jul. 2018, pp. 268–282.

[16] H. Hu, X. Zhang, Z. Salcic, L. Sun, K.-K.-R. Choo, and G. Dobbie, "Source inference attacks: Beyond membership inference attacks in federated learning," *IEEE Trans. Dependable Secure Comput.*, early access, pp. 1–18, 2024.

[17] G. Jignesh Chowdary, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Face mask detection using transfer learning of inceptionv3," in *Proc. Int. Conf. Big Data Analytics*, 2020, pp. 81–90.

[18] *General Data Protection Regulation (GDPR)*. Accessed: Jul. 28, 2024. [Online]. Available: https://gdpr-info.eu/

[19] *California Consumer Privacy Act (CCPA)*. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[20] G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, "VerifyNet: Secure and verifiable federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 911–926, 2020.

[21] C. Wu et al., "It's all in the touch: Authenticating users with HOST gestures on multi-touch screen devices," *IEEE Trans. Mobile Comput.*, vol. 15, pp. 1–14, 2024.

[22] C. Wu, K. He, J. Chen, and R. Du, "ICAuth: Implicit and continuous authentication when the screen is awake," in *Proc. IEEE ICC*, 2019, pp. 1–6.

[23] G. Xu, H. Li, Y. Zhang, S. Xu, J. Ning, and R. H. Deng, "Privacy-preserving federated deep learning with irregular users," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 2, pp. 1364–1381, Mar. 2022.

[24] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Comput. Surveys*, vol. 56, no. 4, pp. 1–34, Apr. 2024.

[25] Y. Wang, Q. Gu, and D. Brown, "Differentially private hypothesis transfer learning," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2019, pp. 811–826.

[26] D. Gao, Y. Liu, A. Huang, C. Ju, H. Yu, and Q. Yang, "Privacy-preserving heterogeneous federated transfer learning," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019.

[27] Y. Zou, Z. Zhang, M. Backes, and Y. Zhang, "Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning," 2020, *arXiv:2009.04872*.

[28] S. Hidano, T. Murakami, and Y. Kawamoto, "TransMIA: Membership inference attacks using transfer shadow training," in *Proc. IEEE IJCNN*, 2021, pp. 1–10.

[29] Y. Long, V. Bindschaedler, and C. A. Gunter, "Towards measuring membership privacy," 2017, *arXiv:1712.09136*.

[30] L. Song, R. Shokri, and P. Mittal, "Privacy risks of securing machine learning models against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 241–257.

[31] Z. Ying, Y. Zhang, and X. Liu, "Privacy-preserving in defending against membership inference attacks," in *Proc. Workshop Privacy-Preserving Mach. Learn. Pract.*, Nov. 2020, pp. 61–63.

[32] S. Pei Liew and T. Takahashi, "FaceLeaks: Inference attacks against transfer learning models via black-box queries," 2020, *arXiv:2010.14023*.

[33] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "MemGuard: Defending against black-box membership inference attacks via adversarial examples," in *Proc. ACM CCS*, 2019, pp. 259–274.

[34] X. Lu et al., "Reinforcement learning-based physical cross-layer security and privacy in 6G," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 425–466, 1st Quart., 2023.

[35] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du, "Liveness is not enough: Enhancing fingerprint authentication with behavioral biometrics to defeat puppet attacks," in *Proc. USENIX Secur.*, 2020, pp. 2219–2236.

[36] C. Wu, J. Chen, K. He, Z. Zhao, R. Du, and C. Zhang, "EchoHand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices," in *Proc. ACM CCS*, 2022, pp. 2931–2945.

[37] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 1897–1914.

[38] K. Patel, "Credit card analytics: A review of fraud detection and risk assessment techniques," *Int. J. Comput. Trends Technol.*, vol. 71, no. 10, pp. 69–79, Oct. 2023.

[39] A. A. S. Alsuwailem, E. Salem, and A. K. J. Saudagar, "Performance of different machine learning algorithms in detecting financial fraud," *Comput. Econ.*, vol. 62, no. 4, pp. 1631–1667, Dec. 2023.

[40] L. Lin et al., "FedLPPA: Learning personalized prompt and aggregation for federated weakly-supervised medical image segmentation," 2024, *arXiv:2402.17502*.

[41] W. Liu, X. Mao, X. Zhang, and X. Zhang, "Efficient sparse least absolute deviation regression with differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2328–2339, 2024.

[42] Z. Sitová et al., "HMOG: New behavioral biometric features for continuous authentication of smartphone users," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 5, pp. 877–892, May 2016.

[43] N. Cariello, S. Levine, G. Zhou, B. Hoplight, P. Gasti, and K. S. Balagani, "SMARTCOPE: Smartphone change of possession evaluation for continuous authentication," *Pervas. Mobile Comput.*, vol. 97, Jan. 2024, Art. no. 101873.

[44] M. Cardaioli, M. Conti, K. Balagani, and P. Gasti, "Your PIN sounds good! Augmentation of pin guessing strategies via audio leakage," in *Proc. ESORICS*, 2020, pp. 720–735.

[45] L. Hu et al., "Defenses to membership inference attacks: A survey," *ACM Comput. Surveys*, vol. 56, no. 4, pp. 1–34, Apr. 2024.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[47] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf

[48] *102 Category Flower Dataset*. Accessed: Jul. 28, 2024. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/flowers/102/

[49] *Dogs Vs. Cats*. Accessed: Jul. 28, 2024. [Online]. Available: https://www.kaggle.com/c/dogs-vs-cats

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[54] C. Wu, K. He, J. Chen, R. Du, and Y. Xiang, "CaIAuth: Context-aware implicit authentication when the screen is awake," *IEEE Internet Things J.*, vol. 7, no. 12, pp. 11420–11430, Dec. 2020.

[55] Z. Chen, H. Li, M. Hao, and G. Xu, "Enhanced mixup training: A defense method against membership inference attack," in *Proc. Int. Conf. Inf. Secur. Pract. Exper.*, 2021, pp. 32–45.

[56] H. Chen et al., "Practical membership inference attack against collaborative inference in industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 477–487, Jan. 2022.

[57] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *Proc. USENIX Secur.*, 2021, pp. 1–18.

[58] Y. Long et al., "A pragmatic approach to membership inferences on machine learning models," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Sep. 2020, pp. 521–534.

[59] B. Hui, Y. Yang, H. Yuan, P. Burlina, N. Z. Gong, and Y. Cao, "Practical blind membership inference attack via differential comparisons," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2021.

[60] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2021, pp. 880–895.

[61] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, "Membership inference attacks by exploiting loss trajectory," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2022, pp. 2085–2098.

[62] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 739–753.

[63] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans, "Revisiting membership inference under realistic assumptions," in *Proc. Privacy Enhancing Technol. Symp.*, 2021.

[64] A. Luqman, A. Chattopadhyay, and K.-Y. Lam, "Membership inference vulnerabilities in peer-to-peer federated learning," in *Proc. Secure Trustworthy Deep Learn. Syst. Workshop*, Jul. 2023, pp. 1–5.

[65] R. Schuster, T. Schuster, Y. Meri, and V. Shmatikov, "Humpty dumpty: Controlling word meanings via corpus poisoning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 1295–1313.

[66] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 2041–2055.

[67] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, "Membership privacy in MicroRNA-based studies," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 319–330.

[68] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[69] D. Lin, X. An, and J. Zhang, "Double-bootstrapping source data selection for instance-based transfer learning," *Pattern Recognit. Lett.*, vol. 34, no. 11, pp. 1279–1285, Aug. 2013.

[70] P.-Y. Jeng, C.-S. Wei, T.-P. Jung, and L.-C. Wang, "Low-dimensional subject representation-based transfer learning in EEG decoding," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 1915–1925, Jun. 2021.

[71] D. Wang, Y. Li, Y. Lin, and Y. Zhuang, "Relational knowledge transfer for zero-shot learning," in *Proc. AAAI*, 2016, pp. 1–7.

[72] Y. Tian, F. Suya, A. Suri, F. Xu, and D. Evans, "Manipulating transfer learning for property inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15975–15984.

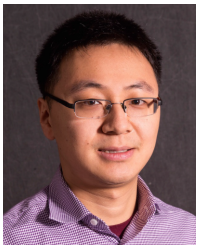
[73] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du, "Toward robust detection of puppet attacks via characterizing fingertip-touch behaviors," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 6, pp. 4002–4018, Nov. 2022.

Cong Wu received the Ph.D. degree from the School of Cyber Science and Engineering, Wuhan University, in 2022. He is currently a Research Fellow with the Cyber Security Laboratory, Nanyang Technological University, Singapore. His leading research outcomes have appeared in USENIX Security, ACM CCS, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON MOBILE COMPUTING. His research interests include biometric security, system security, mobile security, and web3 security.

Jing Chen received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan. He is currently a Full Professor with the School of Cyber Science and Engineering, Wuhan University. He has published more than 120 research papers in many international journals and conferences, including IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON SERVICES COMPUTING, and INFOCOM. His research interests include the areas of network security, cloud security, and mobile security.

Qianru Fang received the bachelor's degree in engineering from Zhejiang University in 2017, and the master's degree in cyberspace security from the School of Cyber Science and Engineering, Wuhan University, in 2022. Her current research interest is trustworthy deep learning.

Kun He (Member, IEEE) received the Ph.D. degree in computer science from the Computer School, Wuhan University. He is currently an Associate Professor with Wuhan University. He has published more than 30 research papers in many international journals and conferences, such as USENIX Security, CCS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING. His research interests include cryptography, network security, cloud security, AI security, and blockchain.



Ziming Zhao received the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, USA, in 2014. He is an Associate Professor with the Khoury College of Computer Sciences and the Director of the CyberspACe securiTy and forensIcs Laboratory (CactiLab), Northeastern University, Boston, USA. His research has been supported by the U.S. National Science Foundation (NSF), the U.S. Department of Defense, the U.S. Air Force Office of Scientific Research, and the U.S. National Centers of Academic Excellence in Cybersecurity (part of the National Security Agency). His research outcomes have appeared in IEEE SECURITY AND PRIVACY, USENIX Security, ACM CCS, NDSS, ACM MobiSys, ACM/IEEE DAC, IEEE RTAS, ACM TISSEC/TOPS, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. His current research interests include systems and software security, network and web security, and human-centric security. He was a recipient of the NSF CAREER Award and the NSF CRII Award. He was also a recipient of the Test-of-Time Paper Award at ACM SACMAT 2024. Additionally, he has received Best/Distinguished Paper Awards from several prestigious conferences, including USENIX Security 2019, ACM AsiaCCS 2022, ACM CODASPY 2014, and ITU Kaleidoscope 2016.

Hao Ren received the Ph.D. degree from the University of Electronic Science and Technology of China in December 2020. He is currently a Research Associate Professor with Sichuan University. He was a Research Fellow with Nanyang Technological University, from July 2022 to February 2024, and with The Hong Kong Polytechnic University, from August 2021 to June 2022. He was a Visiting Ph.D. Student with the University of Waterloo, from December 2018 to January 2020. His research outcomes appeared in major conferences and journals, including WWW, ACM ASIACCS, ACSAC, IEEE TRANSACTIONS ON CLOUD COMPUTING, and IEEE NETWORK. His research interests include data security and privacy, applied cryptography, and privacy-preserving machine learning. He won the Best Paper Award from IEEE BigDataSecurity 2023.

Guowen Xu received the Ph.D. degree from the University of Electronic Science and Technology of China in 2020. He is currently a Post-Doctoral Researcher with the City University of Hong Kong, under the supervision of Prof. Yuguang Fang. He was a Research Fellow with Nanyang Technological University, from March 2021 to May 2023. His research interests include applied cryptography and privacy-preserving deep learning.

Yang Liu (Senior Member, IEEE) is currently a Full Professor and the Director of the Cyber Security Laboratory, Nanyang Technological University, Singapore. He specializes in software security, verification, software engineering, and artificial intelligence. His research has bridged the gap between the theory and practical usage of formal methods and program analysis to evaluate the design and implementation of software for high assurance and security. His work led to the development of state-of-the-art model checker and process analysis toolkit (PAT). He has more than 200 publications and six best paper awards in top tier conferences and journals. With more than 50 million Singapore dollar funding support, he is leading a large research team working on the state-of-the-art software engineering and cyber security problems.

Yang Xiang (Fellow, IEEE) received the Ph.D. degree in computer science from Deakin University, Australia. He is the Dean of digital research and innovation capability platform with the Swinburne University of Technology, Australia. In particular, he is currently leading his team developing active defense systems against large-scale distributed network attacks. He is the chief investigator of several projects in network and system security, funded by the Australian Research Council (ARC). He has published more than 200 research papers in many international journals and conferences, such as IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON INFORMATION SECURITY AND FORENSICS, and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. His research interests include cyber security, which covers network and system security, data analytics, distributed systems, and networking.