Learning Privacy-Preserving Student Networks via Discriminative-Generative Distillation

Shiming Ge, *Senior Member, IEEE,* Bochao Liu, Pengju Wang, Yong Li, and Dan Zeng *Senior Member, IEEE*

***Abstract*—While deep models have proved successful in learning rich knowledge from massive well-annotated data, they may pose a privacy leakage risk in practical deployment. It is necessary to find an effective trade-off between high utility and strong privacy. In this work, we propose a discriminative-generative distillation approach to learn privacy-preserving deep models. Our key idea is taking models as bridge to distill knowledge from private data and then transfer it to learn a student network via two streams. First, discriminative stream trains a baseline classifier on private data and an ensemble of teachers on multiple disjoint private subsets, respectively. Then, generative stream takes the classifier as a fixed discriminator and trains a generator in a data-free manner. After that, the generator is used to generate massive synthetic data which are further applied to train a variational autoencoder (VAE). Among these synthetic data, a few of them are fed into the teacher ensemble to query labels via differentially private aggregation, while most of them are embedded to the trained VAE for reconstructing synthetic data. Finally, a semi-supervised student learning is performed to simultaneously handle two tasks: knowledge transfer from the teachers with distillation on few privately labeled synthetic data, and knowledge enhancement with tangent-normal adversarial regularization on many triples of reconstructed synthetic data. In this way, our approach can control query cost over private data and mitigate accuracy degradation in a unified manner, leading to a privacy-preserving student model. Extensive experiments and analysis clearly show the effectiveness of the proposed approach.**

I. Introduction

DEEP learning [1] has delivered impressive performance in image recognition [2]–[6] due to the powerful capacity of deep networks on learning rich knowledge from large-scale annotated data. However, the deployment of deep models may suffer from the leakage risk of data privacy. Recent works [7], [8] have shown that the private information in the training data can be easily recovered with a few access to the released model. Thus, many real-world requirements [9], [10] need to provide high-performance models while protecting data privacy. Thus, it is necessary to explore a feasible solution that can address a key challenge for model deployment: how to effectively learn a privacy-preserving deep model without remarkable loss of inference accuracy?

Shiming Ge, Bochao Liu, Pengju Wang and Yong Li are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China, and with School of Cyber Security at University of Chinese Academy of Sciences, Beijing 100049, China. Email: {geshiming, liubochao, wangpengju, liyong}@iie.ac.cn.

Dan Zeng is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. E-mail: dzeng@shu.edu.cn. Email: dzeng@shu.edu.cn.

Y. Li is the corresponding author. (e-mail: liyong@iie.ac.cn)

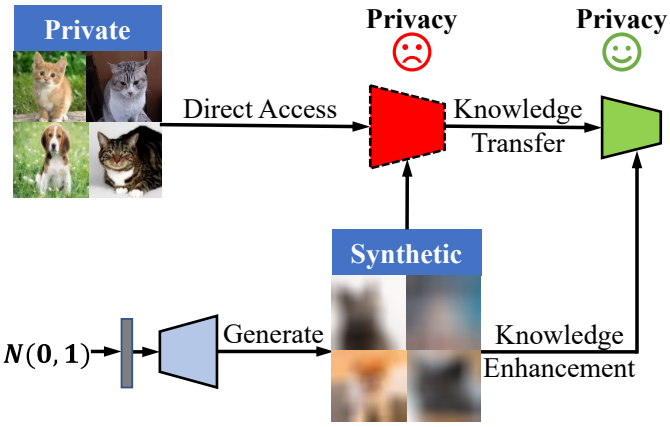


Fig. 1: The models that are learned with direct access to private data may leak privacy. To address that, we aim to learn privacy-preserving models by knowledge transfer from directly-trained discriminative model(s), combined with knowledge enhancement using synthetic data generated by generative model.

Compared to traditional learning solutions that directly access to private data and lead to privacy leakage in released model (in red in Fig. 1), the privacy-preserving learning solutions usually add privacy protection strategies or avoid released model (in green in Fig. 1) directly access to private data during training. Towards this end, many existing approaches have been proposed, which are mainly based on differential privacy [11]. According to the privacy-preserving strategy, they can be roughly grouped into two categories: the implicit category and explicit category.

The "implicit" approaches leverage differentially private learning to train models from private data by enforcing differential privacy on the weights or gradients during training. The prior approach is introducing differential privacy into stochastic gradient descent to train deep networks [12]. Later, Chen *et al.* [13] proposed a stochastic variant of classic backtracking line search algorithm to reduce privacy loss. Papernot *et al.* [14] proposed to replace the unbounded ReLU activation function with a bounded tempered sigmoid function to retain more gradient information. Some recent works proposed gradient operations for private learning by denoising [15], clipping [16], perturbation [17] or compression [18]. Typically, these approaches have promising privacy protection, but often suffer from a big drop in inference accuracy over their non-private counterparts. By contrast, the "explicit" approaches pretrain models on private data and then use auxiliary public or synthetic data to learn the released models with knowledge

transfer by enforcing differential privacy on the outputs of pretrained models [19]–[22]. Papernot *et al.* [19] proposed private aggregation of teacher ensembles (PATE) to learn a student model by privately transferring the teacher knowledge with public data, while Hamm *et al.* [20] proposed a new risk weighted by class probabilities estimated from the ensemble to reduce the sensitivity of majority voting. Zhu *et al.* [22] proposed a practically data-efficient scheme based on private release of k-nearest neighbor queries, which can avoid the decline of accuracy caused by partitioning training set. These approaches generally can improve the model performance while massive unlabeled public data with the same distribution as private data are available. However, these public data are difficult to obtain and the models trained in this way are still at risk of malicious attacks. Recent works [23], [24] trained a private generator and used the generated synthetic data to replace the auxiliary public data. Generally speaking, the most important process in the explicit category is transferring sufficient knowledge from private data to auxiliary data with minimal privacy leakage. Thus, the key issue that needs to be carefully addressed is applying *reliable models* to extract knowledge from private data and exploring *effective auxiliary data* to transfer knowledge.

Inspired by this fact, we propose a teacher-student learning approach to train privacy-preserving student networks via discriminative-generative distillation, which applies *discriminative and generative models* to distill private knowledge and then explores *generated synthetic data* to perform knowledge transfer (Fig. 1). The objective is to enable an effective learning that achieves a promising trade-off between high model utility and strong privacy protection. As shown in Fig. 2, the student is trained by using two streams. First, discriminative stream trains a baseline classifier on all private data and an ensemble of separate teachers on disjoint subsets, while generative stream takes the baseline classifier as a fixed discriminator and trains a generator in a data-free manner. Massive synthetic data are then generated with the generator and used to train a variational autoencoder (VAE) [25]. After that, a few of the synthetic data are fed into the teacher ensemble to query labels with Laplacian aggregation, while most of the synthetic data are fed into VAE to achieve massive data triples by perturbing the latent codes. Finally, a semi-supervised learning is performed by simultaneously handling two tasks: knowledge transfer via supervised data classification, and knowledge enhancement via self-supervised model regularization.

In summary, our approach can effectively learn privacy-preserving student networks by three key components. First, data-free generator learning is incorporated to generate massive synthetic data. These synthetic data are difficult to be identified from appearance but have similar distribution with private data in discriminative space. Therefore, the student learning does not involve any private data and the synthetic data do not expose the information of private data even if they are recovered. Second, differential privacy is incorporated to provide a strong privacy guarantee theoretically. In Laplacian aggregation of teacher ensemble, student's access to its teachers is limited by reducing label queries, so that the student's exposure to teachers' knowledge can be meaningfully quantified and bounded. Third, tangent-normal adversarial regularization is adopted to improve the capacity and robustness of student. In semi-supervised student learning, synthetic data are embedded into the pretrained VAE space and reconstructed from latent codes by adding perturbation along both tangent and normal directions of distribution manifold. Then, the tangent regularization can enforce the local smoothness of the student along the underlying manifold and improve model accuracy, while the normal regularization imposes robustness on the student against noise. In this way, the two regularization terms complement each other, jointly facilitating knowledge transfer from the teacher ensemble to the student. Our approach provides a unified framework to learn privacy-preserving student networks. The data-free generator learning and Laplacian aggregation can protect the private data, and adversarial regularization via VAE reconstruction of the synthetic data can better learn data manifold. Combining them together can protect private data while reducing the impact of noisy labels and instability of the synthetic data.

Our major contributions are three folds: 1) we propose a discriminative-generative distillation approach to train privacy-preserving student networks that achieves an effective trade-off between high utility and strong privacy, 2) we propose to combine data-free generator learning and VAE-based model regularization which facilitates knowledge transfer in a semi-supervised manner, and 3) we conduct extensive experiments and analysis to demonstrate the effectiveness of our approach.

II. Related Works

The approach we proposed in this paper aims to learn privacy-preserving student networks by distilling knowledge from private data and transferring it to synthetic data. Therefore, we briefly review related works from three aspects, including differentially private learning, learning with synthetic data and teacher-student learning.

A. Differentially Private Learning

Differentially private learning [26] aims to address tasks like healthcare [27] where the data are private and the learning process meets differential privacy requirements. Differential privacy provides a guarantee that two adjacent databases produce statistically indistinguishable results under a reasonable privacy budget.

Previous works [28], [29] considered using differential privacy in machine learning settings. Shokri *et al.* [30] introduced a privacy-preserving distributed stochastic gradient descent (SGD) algorithm which applies to non-convex models. Its privacy bound is decided by the number of model parameters that are related to the representation ability of the model, leading to an inefficient trade-off between privacy and model capacity. Abadi *et al.* [12] provided a stricter bound on the privacy loss induced by a noisy SGD by introducing moments accountant. Papernot *et al.* [19] proposed a general framework named private aggregation of teacher ensembles (PATE) for private training. PATE uses semi-supervised learning to transfer the knowledge of the teacher ensemble to the student by using a

differentially private aggregation. It uses the assumption that the student has access to additional unlabeled data. To reduce erroneous aggregation results, Xiang *et al.* [31] proposed a private consensus protocol by returning only the highest voting results above a threshold in aggregation of teacher ensembles, leading to accuracy improvement under the same privacy level. Gao *et al.* [32] improved PATE to securely and efficiently harness distributed knowledge by using lightweight cryptography, which can achieve strong protection for individual labels. Miyato *et al.* [33] proposed virtual adversarial training to avoid the requirements of label information, which reduces queries to the privacy model and protects data privacy. Jagannathan *et al.* [34] combined Laplacian mechanism with decision trees and proposed a random forest algorithm to protect privacy. The idea of differentially private learning can suggest the usage of data for training models under a certain privacy budget.

B. Learning with Synthetic Data

With the development of generative adversarial networks (GANs) [35], recent works began to use synthetic data in training deep networks. Zhang *et al.* [36] found that the performance of classifiers trained in a semi-supervised manner using synthetic data could not be guaranteed and proposed Bad GAN to preferentially select the generator, which greatly improves the feature matching of GANs. Dumoulin *et al.* [37] proposed to jointly learn a generation network and an inference network using synthetic data generated by generation network, achieving a very competitive performance. Salimans *et al.* [38] presented a variety of architectural features and training procedures, which improves the performance of both classifier and generator. Kumar *et al.* [39] proposed to estimate the tangent space to the data manifold using GANs and employ it to inject invariances into the classifier, which can greatly improve in terms of semantic similarity of the reconstructed samples with the input samples. Luo *et al.* [40] introduced smooth neighbors on teacher graphs, which improves the performance of classifier through the implicit self-ensemble of models. Qi *et al.* [41] presented localized GAN to learn the manifold of real data, which could not only produce diverse image transformations but also deliver superior classification performance. The works [23], [42]–[44] used differentially private stochastic gradient descent (DPSGD) to train GANs, which has been proven effective in generating high-dimensional sanitized data [43]. However, DPSGD relies on carefully tuning of the clipping bound of gradient norm, *i.e.*, the sensitivity value. Specifically, the optimal clipping bound varies greatly with model architecture and training dynamics, making the implementation of DPSGD difficult. In order to solve this problem, Chen *et al.* [45] used Wasserstein GANs [46], [47] for a precise estimation of the sensitivity value, avoiding the intensive search of hyper-parameters while reducing the clipping bias. Generally, these approaches aim to generate synthetic data to facilitate model learning, while the privacy issue introduced by generated data is less considered.

C. Teacher-Student Learning

Typically, teacher-student learning applies knowledge distillation [48]–[50] to learn a more compact student model by mimicking the behaviors of a complex teacher model. It is used for model compression while hardly degrading the model performance. In the vanilla knowledge distillation, by using the softmax output of the teacher network as soft labels instead of hard class labels, the student model can learn how the teacher network behaves given tasks in a compact form. Since then, many works [51]–[55] had used and improved this training method. Romero *et al.* [51] proposed to add an additional linear projection layer. Tian *et al.* [56] proposed to combine contrastive learning with knowledge distillation. The teacher-student learning manner has been applied in many applications, such as low-resolution face recognition [57], action recognition [58], semantic segmentation [59], data generation [60] and molecular generation [61]. For circumstances when training data for the teacher are unavailable in practical problems such as privacy, Chen *et al.* [62] proposed a data-free knowledge distillation framework. It regards the pretrained teacher networks as a fixed discriminator and trains a generator to synthesize training samples for the student. To protect the privacy of the data, some works utilize structural improvements [20], such as training a collection of teacher models [63]. Recently, the distillation idea is used to control privacy loss [19], [24]. The key issue in learning privacy-preserving models with distillation is to make knowledge transfer adequately and privately.

III. The Approach

A. Problem Formulation

Given a private dataset $\mathcal{D}$, the objective is learning a privacy-preserving student ϕ_s that does not reveal data privacy and has the capacity approximating to the baseline model ϕ_b trained directly on $\mathcal{D}$. To achieve that, we introduce both discriminative and generative models to enforce the knowledge transfer via discriminative-generative distillation with two streams. Discriminative stream partitions $\mathcal{D}$ into n disjoint subsets $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^n$ and learns an ensemble of multiple teachers $\phi_t = \{\phi_{t,i}\}_{i=1}^n$ where $\phi_{t,i}$ is trained on $\mathcal{D}_i$. Generative stream takes ϕ_b as a fixed discriminator and learns a generator ϕ_g to generate massive synthetic data $\hat{\mathcal{D}}$. A VAE $\{\phi_e, \phi_d\}$ is pretrained on synthetic data, where ϕ_e and ϕ_d are the encoder and decoder respectively. The pretrained VAE is used to obtain data distribution information to facilitate model learning. To reduce the privacy budget, only a few of synthetic data $\hat{\mathcal{D}}_s \subset \hat{\mathcal{D}}$ are used to query the teacher ensemble and get the noisy labels $\hat{\mathcal{L}}_s$. The other unlabelled data $\hat{\mathcal{D}}_u = \hat{\mathcal{D}} \setminus \hat{\mathcal{D}}_s$ with $|\hat{\mathcal{D}}_u| \gg |\hat{\mathcal{D}}_s|$ are employed to provide manifold regularization, with the help of VAE. Thus, the student learning can be formulated by minimizing an energy function $\mathbb{E}$:

$$\mathbb{E}(\mathbb{W}_s; \hat{\mathcal{D}}) = \mathbb{E}_s(\phi_s(\mathbb{W}_s; \hat{\mathcal{D}}_s); \hat{\mathcal{L}}_s) + \mathbb{E}_u(\phi_s(\mathbb{W}_s; \phi_d(\phi_e(\hat{\mathcal{D}}_u))), \phi_s(\mathbb{W}_s; \phi_d(\mathbb{P}[\phi_e(\hat{\mathcal{D}}_u)])])), \tag{1}$$

where $\mathbb{W}_s$ is the parameters of student, $\mathbb{P}[\,]$ is the perturbation operator, $\mathbb{E}_s$ and $\mathbb{E}_u$ are supervised energy term and unsupervised energy term, respectively.

We can see that the risk of privacy leakage can be effectively suppressed due to the isolation between the released student

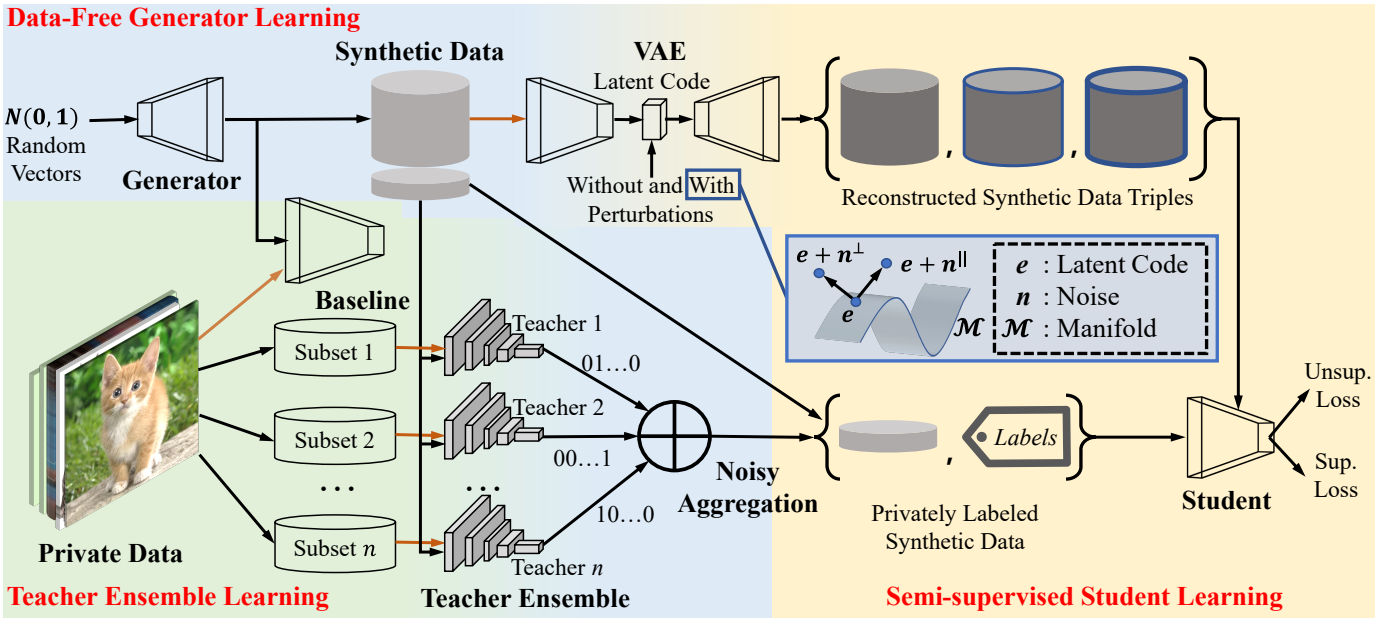


Fig. 2: Overview of the approach. The privacy-preserving student learning is performed with discriminative-generative distillation via two streams. First, discriminative stream trains a baseline classifier on private data and an ensemble of multiple teachers on disjoint private subsets, and generative stream takes the baseline as a fixed discriminator to train a generator in a data-free manner. The generator is employed to generate massive synthetic data that are used to pretrain a variational autoencoder (VAE). Then, the synthetic data are splitted into two parts: a few of them are fed into teacher ensemble in discriminative stream to query labels by noisy aggregation, and most of them are embedded into the VAE space and reconstructed with and without latent code perturbation. Finally, two streams converge to perform semi-supervised student learning by transferring teacher knowledge with few labeled synthetic data and regularizing with massive VAE-reconstructed synthetic data.

model and private data. The supervised energy term can enforce knowledge transfer on the class-related characteristics, while the unsupervised energy term performs self-supervised regularization to enhance knowledge. Towards this end, we solve Eq. (1) via three steps, including: 1) data-free generator learning to get synthetic data $\hat{\mathcal{D}}$ and train a VAE $\{\phi_e, \phi_d\}$, 2) teacher ensemble learning to achieve the labels $\hat{\mathcal{L}}_s$ by differentially private aggregation, and 3) semi-supervised student learning to get $\mathbb{W}_s$.

B. Data-Free Generator Learning

Knowledge transfer from the model trained on private data or synthetic data generated by GAN pretrained on private data may lead to privacy leakage. Meanwhile, the models trained on public data may cause significant accuracy degradation due to distribution mismatch, since finding public data that match the distribution with private data often is very difficult [19]. Moreover, the resulting models are vulnerable to attacks since adversaries can also access public data. Thus, we aim to learn a generator in a data-free manner to generate synthetic data, which does not compromise privacy to assist in knowledge transfer from private data to learn released models.

Unlike traditional GAN training where the discriminator is an online learned two-class classifier, our data-free generator learning first pretrains a baseline multi-class classifier ϕ_b (with parameters $\mathbb{W}_b$) on private data $\mathcal{D}$ that serves as the fixed discriminator, and then train the generator ϕ_g without data. It is suggested that the tasks of discrimination and classification can improve each other and the multi-class classifier can learn the data distribution better than the two-class discriminator [62], [64]. Thus, the key of using the multi-class classifier as discriminator is defining a loss to evaluate the generated data. Towards this end, we assess a synthetic example $\mathbf{x} = \phi_g(\mathbb{W}_g; \mathbf{z})$ generated by ϕ_g with parameters $\mathbb{W}_g$ from a random vector $\mathbf{z}$ by the following loss:

$$\begin{aligned}\mathcal{L}(\mathbf{x}) =& \ell(\phi_b(\mathbb{W}_b; \mathbf{x}), \arg\max_j(\phi_b(\mathbb{W}_b; \mathbf{x}))_j) + \\ & \alpha\phi_b(\mathbb{W}_b; \mathbf{x}) \log \phi_b(\mathbb{W}_b; \mathbf{x}) + \beta\|\phi_b(\mathbb{W}_b^-; \mathbf{x})\|_1,\end{aligned} \tag{2}$$

where α and β are the tuning parameters to balance the effect of three terms, and we set them as 5 and 0.1 respectively. The first term $\ell(.)$ is cross entropy function that measures the one-hot classification loss, which enforces the generated data having similar distribution as the private data. The second term is the information entropy loss to measure the class balance of generated data. The third term uses l_1-norm $\|*\|_1$ to measure the activation loss, since the features $\phi_b(\mathbb{W}_b^-; \mathbf{x})$ that are extracted by the discriminator and correspond to the output before the fully-connected layer tend to receive higher activation value if input data are real rather than some random vectors, where $\mathbb{W}_b^- \subset \mathbb{W}_b$ is the discriminator's backbone parameters. Then, using the fixed discriminator, the generator learning is performed iteratively via five steps:

- randomly generate a batch of noise vectors : $\{\mathbf{z}_i\}_{i=1}^m$.

- generate synthetic samples $\{\mathbf{x}_i\}_{i=1}^m$ for training: $\mathbf{x}_i = \phi_g(\mathbb{W}_g; \mathbf{z}_i)$.
- apply the discriminator on the mini-batch: $\mathbf{y}_i = \phi_b(\mathbb{W}_b; \mathbf{x}_i)$.
- calculate the loss function with Eq. (2) on mini-batch: $\sum_i \mathcal{L}(\mathbf{x}_i)$.
- update weights $\mathbb{W}_g$ using back-propagation.

In this way, the synthetic data $\hat{\mathcal{D}}$ generated by the learned generator have a similar distribution to private data without compromising privacy. Fig. 5 shows some examples. The synthetic data are very helpful for student learning, which can greatly improve accuracy compared to using public data and reduce accuracy loss compared to using private data directly. With $\hat{\mathcal{D}}$, we train a VAE $\{\phi_e, \phi_d\}$, where the encoder ϕ_e with parameters $\mathbb{W}_e$ and decoder ϕ_d with parameters $\mathbb{W}_d$ are constructed with convolutional neural networks like [65].

C. Teacher Ensemble Learning

Instead of using a single model as teacher that may lead to privacy leakage [63], we learn an ensemble of teachers for knowledge transfer. Towards this end, we partition the private data $\mathcal{D}$ into n disjoint subsets $\{\mathcal{D}_i\}_{i=1}^n$ and then separately train a teacher $\phi_{t,i}$ with parameters $\mathbb{W}_{t,i}$ on each subset $\mathcal{D}_i$, leading to the teacher ensemble $\phi_t = \{\phi_{t,i}\}_{i=1}^n$.

In general, the number of teachers n has an impact on knowledge extraction from private data. When n is too large, the amount of each training subset data gets less and the teachers may be underfitted. When n is too small, it will make the noise of differential privacy more influential and lead to unusable aggregated labels. Thus, the teacher number n should be carefully set in experience.

The teacher ensemble serves to label query, where the synthetic data $\mathbf{x} \in \hat{\mathcal{D}}_s$ is fed and the predicted labels by multiple teachers are privately aggregated:

$$l = \arg\max_k \{\mathcal{V}_k(\{\phi_{t,i}(\mathbb{W}_{t,i}; \mathbf{x})\}_{i=1}^n) + Lap(2/\varepsilon_0)\}, \quad (3)$$

where $\mathcal{V}_k(\cdot)$ counts the votes of the query being predicted as class k by all n teachers, the final predicted label l is noisy and used to supervise the student training, a low privacy budget ε_0 is used to adjust privacy protection and $Lap(2/\varepsilon_0)$ denotes the Laplacian distribution with location 0 and scale $2/\varepsilon_0$. For student training, each example from the query data $\hat{\mathcal{D}}_s$ is fed into the teacher ensemble and then the prediction is privately aggregated via Laplacian aggregation in Eq. (3), leading to $\hat{\mathcal{L}}_s = \{l_i\}_{i=1}^{|\hat{\mathcal{D}}_s|}$. Directly using the maximum value of vote counts as labels may leak privacy, so we add random noise to the voting results to introduce ambiguity. Intuitively, this means that multiple teachers jointly determine the query result, making it difficult for adversary to recover the training data. In addition to this, our approach can provide the same or stronger privacy guarantee than many state-of-the-arts [17]–[19], [23], [24], [45], [66], [67] while reducing accuracy degradation by knowledge enhancement with an extra model regularization. It also means that our approach will have a less privacy cost when delivering student models with the same accuracy.

D. Semi-Supervised Student Learning

To reduce privacy leakage, we only use a few of synthetic data $\hat{\mathcal{D}}_s$ for label query. Thus, the teacher knowledge that transfers from private data to $\hat{\mathcal{D}}_s$ is not only noisy due to Laplacian aggregation but also insufficient due to limited data. To enhance knowledge transfer, we learn the student in a semi-supervised fashion by adding another unsupervised pathway. Each synthetic example $\mathbf{x}_j \in \hat{\mathcal{D}}_u$ is embedded into VAE space and get a mean vector μ_j and a standard deviation vector σ_j with $\{\mu_j, \sigma_j\} = \phi_e(\mathbb{W}_e; \mathbf{x}_j)$ that form a normal distribution $\mathcal{N}(\mu_j, \sigma_j)$. Then, the data is reconstructed from a sampled code $\mathbf{e}_j$ as well as its perturbed versions along tangent and normal directions of the distribution manifold, leading to massive data triples $\mathcal{T} = \{(\hat{\mathbf{x}}_j, \hat{\mathbf{x}}_j^{\parallel}, \hat{\mathbf{x}}_j^{\perp})\}_{j=1}^{|\hat{\mathcal{D}}_u|}$ with:

$$\hat{\mathbf{x}}_j = \phi_d(\mathbb{W}_d; \mathcal{M}(\mathbf{e}_j)),\ \hat{\mathbf{x}}_j^* = \phi_d(\mathbb{W}_d; \mathcal{M}(\mathbf{e}_j + \mathbf{n}_j^*)), \quad (4)$$

where the mapping operator $\mathcal{M}(\cdot)$ projects the code into decoder input, $\mathbf{n}_j^*$ is random perturbation noise along tangent direction ($* = \parallel$) or normal direction ($* = \perp$). Then, the semi-supervised student learning is performed with $\{\hat{\mathcal{D}}_s, \hat{\mathcal{L}}_s\}$ and $\mathcal{T}$. The supervised energy in Eq. (1) can be formulated as

$$\mathbb{E}_s = \sum_{i=1}^{|\hat{\mathcal{D}}_s|} \ell(\phi_s(\mathbb{W}_s; \mathbf{x}_i), l_i),\ s.t.\ \mathbf{x}_i \in \hat{\mathcal{D}}_s,\ l_i \in \hat{\mathcal{L}}_s, \quad (5)$$

and the unsupervised energy is formulated as

$$\begin{aligned}\mathbb{E}_u = &\sum_{j=1}^{|\hat{\mathcal{D}}_u|} \|\phi_s(\mathbb{W}_s^-; \hat{\mathbf{x}}_j) - \phi_s(\mathbb{W}_s^-; \hat{\mathbf{x}}_j^{\perp})\|^2 \\ &+ \sum_{j=1}^{|\hat{\mathcal{D}}_u|} \|\phi_s(\mathbb{W}_s^-; \hat{\mathbf{x}}_j) - \phi_s(\mathbb{W}_s^-; \hat{\mathbf{x}}_j^{\parallel})\|^2 \\ &+ \sum_{j=1}^{|\hat{\mathcal{D}}_u|} \phi_s(\mathbb{W}_s; \hat{\mathbf{x}}_j) \log \phi_s(\mathbb{W}_s; \hat{\mathbf{x}}_j),\end{aligned} \quad (6)$$

where $\mathbb{W}_s^- \subset \mathbb{W}_s$ is backbone parameters of the student for extracting features. We can see that the unsupervised energy Eq. (6) includes normal regularization, tangent regularization and entropy regularization. The first two regularization terms enhance model robustness against perturbations along orthogonal and parallel directions to the underlying data manifold respectively, while entropy regularization ensures the student output more determinate predictions. This tangent-normal adversarial regularization by adding perturbation to the latent layer can make the student vary smoothly along tangent space and have strong robustness along normal space [65].

The total energy comes from two streams. Discriminative stream employs the supervised energy for knowledge transfer from teacher ensemble with a few of query examples, while generative stream takes the unsupervised energy for knowledge enhancement via model regularization. The differentially private aggregation provides privacy protection, while the usage of VAE embedding and reconstruction can obtain the characteristics of the data in the tangent and normal spaces. In particular, generative stream applies self-supervised learning from massive unannotated data to compensate the knowledge that may miss in discriminative stream.

TABLE I: Performance comparisons with 6 explicit approaches: Test accuracy (%) and the drop (∇) with respect to baseline under different privacy budget ε ($\delta = 10^{-5}$, BL: Baseline).

Approach	ε	MNIST (BL: 99.2 ∇)	FMNIST (BL: 91.0 ∇)	ε	MNIST (BL: 99.2 ∇)	FMNIST (BL: 91.0 ∇)
DP-GAN [23]	10.0	60.1 (39.1)	50.9 (40.1)	1.00	40.3 (58.9)	10.5 (80.5)
PATE-GAN [24]	10.0	66.7 (32.5)	62.2 (28.8)	1.00	41.7 (57.5)	42.2 (48.8)
G-PATE [66]	10.0	80.9 (18.3)	69.3 (21.7)	1.00	58.8 (40.4)	58.1 (32.9)
DP-MERF [67]	10.0	68.7 (30.5)	62.5 (28.5)	1.00	65.0 (34.2)	61.0 (30.0)
GS-WGAN [45]	10.0	80.0 (19.2)	65.0 (26.0)	1.00	14.3 (84.9)	16.6 (74.4)
DataLens [18]	10.0	80.7 (18.5)	70.6 (20.4)	1.00	71.2 (28.0)	64.8 (26.2)
Our DGD	10.0	**97.4** (**1.80**)	**88.2** (**2.80**)	1.00	**73.6** (**25.6**)	**64.9** (**26.1**)

E. Discussion

Practical Deployment. To learn a privacy-preserving student, our approach trains it from synthetic data generated with a generator pretrained in a data-free manner. Typically, the learning could be deployed in a single server where the private data are partitioned into several subsets to train an ensemble of teachers. Moreover, the learning is also suitable to deploy for jointly training models from distributed clients via a trusted server as the coordinator. In this case, each client trains a teacher on its private local data and all teachers form the teacher ensemble, while the trusted server aggregates the local data via centered learning or local knowledge via federated learning [68] to pretrain a baseline classifier that is used to train a generator in a data-free manner. Then, the server applies the generator to generate massive synthetic data that are used to pretrain a VAE. After that, the server splits the synthetic data into two parts: a few of them are distributed to local clients to query labels by noisy aggregation in discriminative stream, and most of them are fed into generative stream for VAE reconstruction to get massive synthetic data triples. Finally, the student is trained on noisy labels and synthetic data triples in a semi-supervised manner within the trusted server. By allowing only student to be accessible to adversaries, the trained student could be deployed on practical applications and gives the differential privacy guarantee, introduced next.

Privacy Analysis. According to the learning process in two streams, the total privacy budget contains two parts. Discriminative privacy budget is computed as PATE [19], [21], achieving ε_0-differential privacy via Eq. (3) and getting $(|\hat{\mathcal{D}}_s|\varepsilon_0^2 + \varepsilon_0\sqrt{-2|\hat{\mathcal{D}}_s|\log\delta}, \delta)-$differential privacy over $|\hat{\mathcal{D}}_s|$ queries for all $\delta \in (0, 1)$ [69]. Generative privacy budget is computed according to the latent code perturbation in VAE construction. By taking the synthetic data in generative stream as a sequence, we achieve ε_1-differential privacy by adding Laplacian noise with scale $2c/\varepsilon_1$ to the normalized latent codes, where c is the dimension of latent codes. It could be explained as follow. According to Laplacian mechanism and post-processing theorem [69], we have: for any two different images $\mathbf{x}_j$ and $\mathbf{x}'_j$ as well as possible reconstructed output $\hat{\mathbf{x}}_j$, the VAE reconstruction mechanism $\mathcal{A}$ satisfies $\Pr[\mathcal{A}(\mathbf{x}_j) = \hat{\mathbf{x}}_j] \leq \exp(\varepsilon_1) \cdot \Pr\left[\mathcal{A}\left(\mathbf{x}'_j\right) = \hat{\mathbf{x}}_j\right]$ where $\Pr[\cdot]$ is the probability function. Then, we have the theorem.

Theorem 1 *The sequence of VAE reconstruction mechanism $\mathcal{A}$, denoted as $\mathcal{A}(\hat{\mathcal{D}}_u)$ satisfies ε_1-differential privacy.*

Proof *For any two adjacent datasets $\hat{\mathcal{D}}_u$ and $\hat{\mathcal{D}}'_u$ where $\mathbf{x}_j \in \hat{\mathcal{D}}_u$ and $\mathbf{x}'_j \in \hat{\mathcal{D}}'_u$ are the two only different images, we have*

$$\begin{aligned}
&\Pr\left[\mathcal{A}\left(\hat{\mathcal{D}}_u\right) \subseteq \mathcal{O}\right] \\
&= \Pr\left[\mathcal{A}\left(\hat{\mathcal{D}}_u \cap \hat{\mathcal{D}}'_u\right) \subseteq \mathcal{O}\right] \cdot \Pr\left[\mathcal{A}\left(\mathbf{x}_j\right) = \hat{\mathbf{x}}_j\right] \\
&\leq \exp\left(\varepsilon_1\right) \cdot \Pr\left[\mathcal{A}\left(\hat{\mathcal{D}}_u \cap \hat{\mathcal{D}}'_u\right) \subseteq \mathcal{O}\right] \cdot \Pr\left[\mathcal{A}\left(\mathbf{x}'_j\right) = \hat{\mathbf{x}}_j\right] \\
&= \exp(\varepsilon_1) \cdot \Pr\left[\mathcal{A}\left(\hat{\mathcal{D}}'_u\right) \subseteq \mathcal{O}\right],
\end{aligned} \tag{7}$$

where $\mathcal{O}$ denotes the subset of possible outputs. Eq. (7) indicates that $\mathcal{A}(\hat{\mathcal{D}}_u)$ satisfies ε_1-differential privacy according to the definition of differential privacy [69]. Further, according to the composition theorem [69], our approach finally satisfies $(|\hat{\mathcal{D}}_s|\varepsilon_0^2 + \varepsilon_0\sqrt{-2|\hat{\mathcal{D}}_s|\log\delta} + \varepsilon_1, \delta)$-differential privacy and gives the differential privacy guarantee.

IV. Experiments

To verify the effectiveness of our proposed discriminative-generative distillation approach **DGD**, we conduct experiments on four datasets (MNIST [70], Fashion-MNIST [71] (FMNIST), SVHN [72] and CIFAR-10 [73]) and perform comparisons with 13 state-of-the-art benchmarks, including 6 explicit approaches that train models with generative data (DP-GAN [23], PATE-GAN [24], GS-WGAN [45], G-PATE [66] and DP-MERF [67], DataLens [18]), and 7 implicit approaches that train models with differentially private learning (DPSGD [12], zCDP [74], GEDDP [15], DP-BLSGD [13], RDP [75], TSADP [14] and GEP [17]). Here, all explicit approaches but DP-GAN apply teacher-student learning to distill models, while all implicit approaches perform differentially private learning without model distillation. To make the comparisons fair, our experiments use the same experimental settings as these benchmarks and take the results from their original papers. Note that original PATE [19] conducted experiments with private data to simulate public data, thus we just compare to it in component analysis experiment.

MNIST and FMNIST are both 10-class datasets containing 60K training examples and 10K testing examples. The examples are 28×28 grayscale handwriting digit images or fashion images. SVHN is a real-world 32×32 color digit image dataset that contains 73257 training examples, 26032 testing examples and 531131 extra training examples. CIFAR10 consists of 60K

32×32 color images in 10 classes, including 50K for training and 10K for testing.

For each dataset, we take its training examples as private data and directly learn a baseline classifier as the discriminator as well as an ensemble of teachers, and then transfer the teacher knowledge to learn student. We set the Laplacian noise scale to be $2/\varepsilon_0 = 40$. We generate the same number of synthetic data as the private training data with the learned generator by randomly generating latent codes and then feeding into the generator to require synthetic data, *e.g.*, generating 60K synthetic images for MNIST. In VAE reconstruction, we set $c = 32$ and $\varepsilon_1 = 0.01$. The models are evaluated on testing examples with privacy cost, classification accuracy and accuracy drop with respect to baseline.

We mainly use simple network structures that are the same to the benchmarks for teachers and student to conduct the experiments. On MNIST and FMNIST, the networks of baseline and teachers have the same structure, which contains two 3×3 convolutional layers (with ReLU activation and max-pooling, and 64 and 128 channels, respectively) to extract features hierarchically, followed by the softmax output layer indicating 10 classes. Each convolutional layer has the stride of 1 with 1-padding and are randomly initialized with Xavier. On SVHN, we add two extra fully-connected layers (with 384 and 192 neurons) with ReLU. We use Adam optimization algorithm to learn all models, and set batch size as 128. To learn all teachers, the iteration rounds are 3000, the learning rate is first set to 0.05 and decays linearly with iteration round to 0. For generator learning, the iteration rounds are 200, the learning rate is first set to 0.2 and 10 times decays every 80 rounds. To learn VAE and student, the iteration rounds are 500, the learning rate is first set to 0.001 and decays linearly with iteration round to 0. The teacher number on these three datasets is 250. We also study a complex structure for teachers and conduct an experiment on CIFAR10 with 100 teachers. Here, we fine-tune a vision transformer [6] on CIFAR10 training set and modify the dimension of the last fully-connected layer to 10. The model is pretrained on ImageNet and gives a top-1 classification accuracy of 81.4%.

A. State-of-the-Art Comparison

We conduct comparisons with 6 explicit approaches under different privacy budget on MNIST and FMNIST and 7 implicit approaches on CIFAR10. The performance is evaluated with test accuracy of student and accuracy drop with respect to its baseline under the condition of (ε, δ)-differential privacy. Here, ε is privacy budget and δ is failure probability. A lower privacy budget means a stronger privacy guarantee.

Comparisons with 6 Explicit Approaches. In the comparisons, we check the performance under different privacy budget, and report the results in Tab. I. Our approach takes 1300 queries under $\varepsilon = 10.0$ and 27 queries under $\varepsilon = 1.00$, respectively. All approaches are under a low failure probability $\delta = 10^{-5}$. The test accuracy of baseline model is 99.2% on MNIST and 91.0% on FMNIST.

From Tab. I, under the same condition of high privacy budget, we can see that our student achieves the highest test accuracy of 97.4% on MNIST and 88.2% on FMNIST, which remarkably reduces the accuracy drop by 1.80% and 2.80% respectively. It shows that our approach has the best privacy-preserving ability and minimal accuracy drop. Under the same low privacy budget, all approaches suffer from accuracy drop with respect to their counterparts under high privacy budget. However, our student still delivers the highest test accuracy and the lowest accuracy drop on both datasets. These results imply that discriminative stream plays an important role in knowledge transfer from private data. First, discriminative stream provides class identity supervision thus we cannot just use generative stream. Second, it uses certain queries to balance privacy protection and model accuracy.

TABLE II: Test accuracy comparisons with 7 implicit approaches on CIFAR10 under different ε ($\delta = 10^{-5}$).

Approach	ε	Accuracy (%)
DPSGD [12]	3.19	60.7
zCDP [74]	6.78	44.3
GEDDP [15]	3.00	55.0
DP-BLSGD [13]	8.00	53.0
RGP [75]	8.00	63.4
TSADP [14]	7.53	66.2
GEP [17]	5.00	70.1
Our DGD	**3.00**	**73.6**

Comparisons with 7 Implicit Approaches. In addition, we conduct experimental comparison on CIFAR10 and report in Tab. II, where our approach achieves the highest accuracy of 73.6% under the lowest privacy budget of 3.00. The main reason comes from that our approach adopts an extra generative stream to enhance knowledge transfer with massive synthetic data generated by a data-free learned generator. In this way, the missing knowledge can be recovered from generative stream and the accuracy can be improved.

B. Component Analysis

After the promising performance achieved, we further analyze the impact of each component in our approach, including label query, generator learning, teacher ensemble, noisy aggregation, and VAE-based regularization.

Label Query. To study query effect on the trade-off between model accuracy and privacy protection, we compare the student learning under 27, 750, 1000 and 3000 queries. We treat the label of a generated example by the teacher ensemble as a query. The query number determines the privacy budget and failure probability, and we use differential privacy with moments accountant [19] as metric. More queries will cost a larger privacy budget and fixed query number will lead to constant privacy cost. The results are shown in Fig. 3. They are as expected where a higher privacy budget leads to a higher model accuracy. Besides, in our approach, the private information that the delivered student can directly access is the noisy teachers' prediction outputs who pass through Laplacian aggregation. The results also reveal that our student learning by dicriminative-generative distillation can be performed robustly

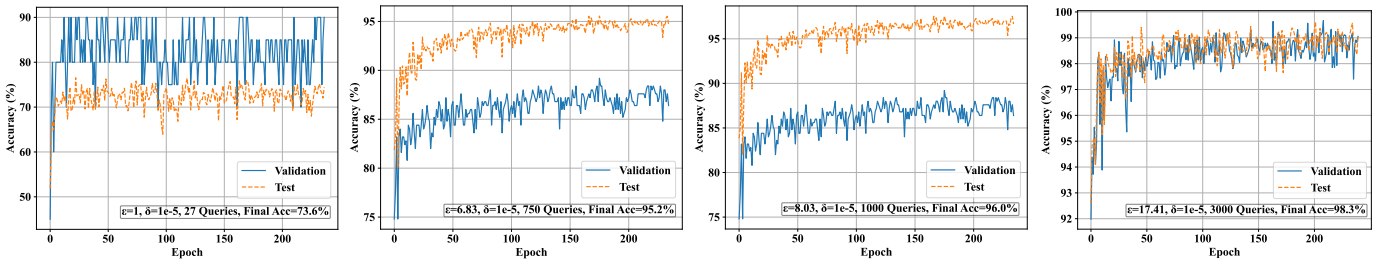


Fig. 3: Model accuracy during training and privacy cost under different queries

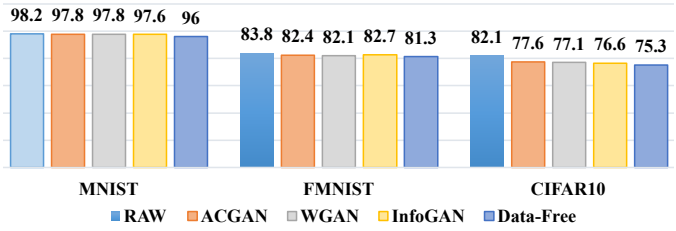


Fig. 4: The effect of different generators on student accuracy (%). Here, RAW means learning with private data.

Fig. 5: The generated images with different generator learning approaches. It is obvious that the examples generated by data-free learned generator protect privacy better.

and consistently under different label queries and providing a certain number of examples (*e.g.*, 750) can lead to an impressive accuracy of 95.2%. It is very helpful in many practical applications where a few of samples are available for sharing. Therefore, our approach can effectively learn privacy-preserving student models and control accuracy drop.

Generator Learning. To study the effect of data generation, we conduct student learning on MNIST, FMNIST and CIFAR10 with the raw private data as well as synthetic data generated by four generative approaches, including ACGAN [76], WGAN [46], InfoGAN [77] and our Data-Free learned generator. The results are shown in Fig. 4 and some generated examples can be seen in Fig. 5. It is easy to distinguish the images generated by ACGAN, WGAN and

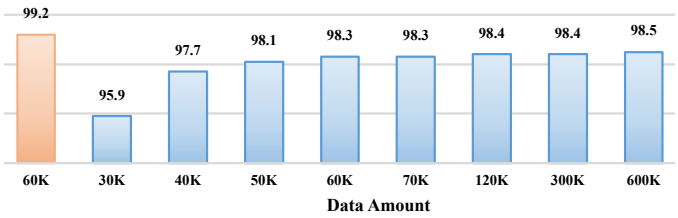


Fig. 6: The accuracy (%) under different amounts of synthetic data, which is generated with the generator trained by taking the baseline (in orange) as the fixed discriminator. The baseline is trained on 60K private training data of MNIST.

InfoGAN, implying that these generators learned with private data may expose data privacy in spite of achieving higher accuracy. By contrast, the synthetic images generated by data-free learned generator are hardly identified by human. Thus, the data-free learned generator effectively protects privacy while delivering comparable accuracy since it matches the distribution of private data in discriminative space.

Beyond the generator learning method, we further check the impact of synthetic data. Towards this end, via the generator trained with the baseline on MNIST as the fixed discriminator, we generate 8 synthetic datasets with various amounts to train students and report their performance in Fig. 6. We can find that the model accuracy increases by training on more synthetic data and gets smooth after the used synthetic data reaches 60K that is equal to the number of private training examples. Therefore, we generate the same number of synthetic data as the private training data to provide a good trade-off between model performance and training efficiency.

Teacher Ensemble. We check the effect of teacher number on the accuracy of teacher ensemble. The top left of Fig. 8 shows the results on three datasets for evaluating the effect on simple and complex classification tasks. We find that the accuracy increases along with teacher number within a certain range, indicating that the model performance can be boosted by increasing teacher number in a certain range. It is very helpful in real-world applications like federated learning [68], [78] where the private model can be improved by adding the sharing data parties. The performance starts to degrade when the teacher number increases to a certain value. Then the amount of partitioned training data for each teacher starts to become inadequate for learning, suggesting careful selection of the teacher number.

VAE-based Regularization. To check the effect of VAE on

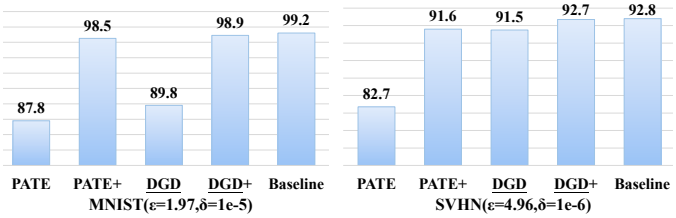


Fig. 7: The effect of VAE on accuracy (%) under different privacy budget and private aggregation mechanism.

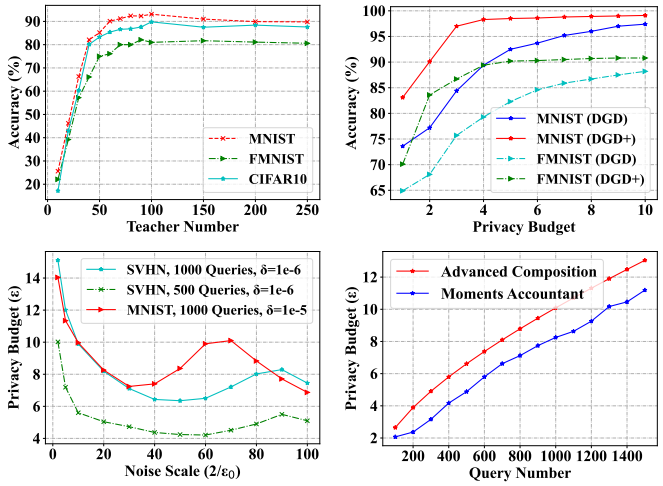


Fig. 8: Top left: The privately aggregated accuracy of teacher ensemble under different teacher numbers. Top right: The student accuracy under different privacy budget ε ($\delta = 10^{-5}$). Bottom left: Privacy budget under different noise scale. Bottom right: Privacy budget under different query.

student learning, we modify our approach for comparing to PATE [19] with Laplacian aggregation as well as its improved variant PATE+ [21] with Gaussian aggregation under the same experimental settings. Towards this end, we remove the generator and feed private training data (simulating public unlabeled data like [19], [21]) into VAE to learn students with Laplacian and Gaussian aggregation, leading to two modified approaches denoted as DGD and DGD+, respectively. We train various students on MNIST and SVHN under different privacy budget and conduct the comparisons. In our experiments, a few of training data serve as queries in noisy aggregation and the remaining most of the training data are fed into VAE where each example is reconstructed into a synthetic triple. The results are reported in Fig. 7, where using VAE in our DGD and DGD+ can consistently improve model accuracy over PATE and PATE+ without sacrificing privacy guarantee, respectively. For example, DGD+ delivers an accuracy of 92.7% on SVHN that is very close to 92.8% achieved with baseline, implying the effectiveness of VAE-based regularization, since it can provide self-supervised knowledge enhancement to compensate the accuracy drop. We also achieve higher accuracy by Gaussian aggregation than Laplacian aggregation (*i.e.*, PATE+ vs. PATE, and DGD+ vs. DGD) as stated in [21], implying the significance of noisy aggregation, introduced next.

Noisy Aggregation. During the student learning, the obtained query labels are disturbed by noise with a scale of $2/\varepsilon_0$. Theoretically, the higher the noise scale is, the lower privacy cost and better privacy protection is. However, too high noise scale may cause label distortion, making student difficult to learn useful knowledge and unsuitable for practical deployment. We study how the noise scale $2/\varepsilon_0$ affects the privacy cost and report in the bottom left of Fig. 8. We can observe that the privacy cost declines rapidly with the noise scale within a certain range, has a short rise in the middle, and then keeps smooth. We suspect the main reason is the calculation process of privacy protection metric with moments accountant. Thus, we could select a noise scale (*e.g.*, $2/\varepsilon_0 = 30$) to provide a good trade-off between privacy protection and model accuracy.

To further study the effect of noisy aggregation, we conduct experiments on MNIST and FMNIST and report the results in the top right of Fig. 8 where DGD+ improves DGD with Gaussian aggregation. It shows that the improvement of all students' accuracy starts rapid and gets smooth when increasing the privacy budget. Moreover, as expected, the accuracy of students trained with Gaussian aggregation is remarkably improved over Laplacian aggregation, suggesting that more advanced noisy aggregation mechanisms can be incorporated into our framework to facilitate performance.

C. Privacy-Preserving Analysis

We further study the privacy-preserving ability of the learned students. Towards this end, we conduct both theoretical analysis and practical analysis.

In theory, our approach trains students from VAE-reconstructed synthetic data in generative stream and noisy labels in discriminative stream, where the reconstructed synthetic data are achieved by inputting synthetic data into VAE and reconstructing from noisy latent codes. As discussed in III-E, it contains two parts of privacy budgets and totally achieves $(|\hat{\mathcal{D}}_s|\varepsilon_0^2 + \varepsilon_0\sqrt{-2|\hat{\mathcal{D}}_s|\log\delta} + \varepsilon_1, \delta)$−differential privacy over $|\hat{\mathcal{D}}_s|$ queries for all $\delta \in (0, 1)$. In our experiments, generative privacy budget $\varepsilon_1 = 0.01$ is very small and can be ignored in the total privacy budget. Discriminative privacy budget dominates the total privacy budget, *e.g.*, having a much higher privacy budget of 5.80 under $\varepsilon_0 = 1/20$, $\delta = 10^{-5}$ over 400 queries. For discriminative stream, we first use differential privacy with advanced composition [69] to track privacy loss, seeing the red curve in the bottom right of Fig. 8. To better track the privacy loss, we further use differential privacy with moment accountant [12] and add an advanced limit [19] to get a lower privacy budget. The results can be seen the blue curve in the bottom right of Fig. 8. We can see that the privacy guarantee is satisfied in all metrics, *e.g.*, a privacy budget of 10.1 with advanced composition or 8.03 with moments accountant under 1000 queries, leading to an effective trade-off between privacy-preserving model learning and accuracy drop control.

In practice, a learning process of our approach can achieve several major models, including: 1) the baseline model (serving as the fixed discriminator) and the ensemble of teachers that are kept privately, 2) the data-free learned generator that

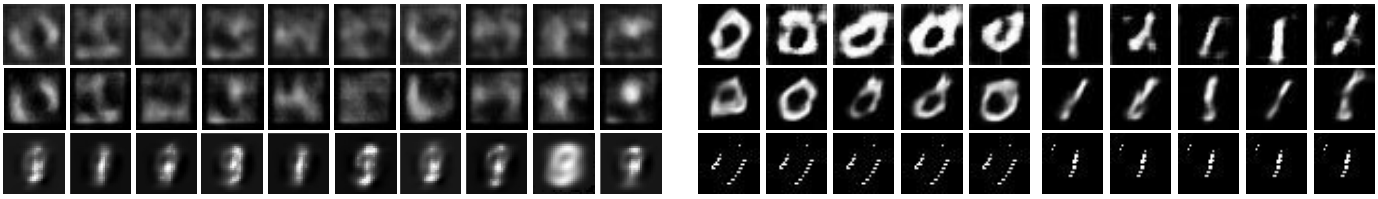


Fig. 9: The left shows some reconstructed images with different reconstruction attacks [62] (Top), [7] (Middle) and [8] (Bottom) on our student learned on MNIST. The right gives the results of model inversion attack [8] against binary classification models trained on the subset of MNIST only containing 0 and 1 with DataLens (Top), GS-WGAN (Middle) and DGD (Bottom).

can be delivered to provide valuable synthetic data for training more further models in a more privacy-preserving manner than using private or other generative data (as shown in Fig. 5 and discussed above), and 3) the student that is delivered for privacy-preserving deployment. We take the student learned on MNIST under $(10.0, 10^{-5})$-differential privacy as an example and study its privacy-preserving ability against three reconstruction attacks, including reconstruction with data-free learned generator [62], model inversion attack with confidence information and basic countermeasures [7], and adversarial model inversion attack with background knowledge alignment [8]. Some reconstructed results by these three attacks are shown in the left of Fig. 9, from the first to the third row, respectively. We can see that the reconstructed images are very different from the original images and hardly distinguished which number they are by human. Thus, the student can well protect data privacy whilst delivering a high accuracy of 97.4%. To further verify the privacy-preserving ability of our students, we investigate an especial case by conducting inversion attack [8] against binary classification models that are trained on a subset of MNIST containing only '0's and '1's with GS-WGAN, DataLens and our DGD. The results are shown in the right of Fig. 9. The reconstruction results of GS-WGAN and DataLens can be distinguished by human, while our model can provide better protection against reconstruction attacks. From these results, we can safely claim that our approach can provide effective privacy-preserving model learning and the learned models are suitable particularly for practical application on privacy-conscious scenarios.

V. Conclusion

Deep models trained on private data may pose the risk of privacy leakage. To facilitate model deployment, we proposed a dicriminative-generative distillation approach to learn privacy-preserving student networks. The approach takes dicriminative and generative models as bridge to distill knowledge from private data and transfer it to learn students in a semi-supervised manner. The supervised learning from noisy aggregation of multiple teachers can provide privacy guarantee, while the unsupervised learning from massive synthetic generated by a data-free learned generator can reduce accuracy drop. Extensive experiments and analysis were conducted to show the effectiveness of our approach. In the future, we will devise more advanced differential privacy mechanisms to improve the approach and explore the approach in more real-world applications like federated learning on medical images.

Acknowledgements. This work was partially supported by grants from the Beijing Natural Science Foundation (19L2040), National Key Research and Development Plan (2020AAA0140001), and National Natural Science Foundation of China (61772513).

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[7] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 1322–1333.

[8] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019, pp. 225–240.

[9] J. M. Abowd, "The u.s. census bureau adopts differential privacy," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018, pp. 2867–2867.

[10] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2014, pp. 1054–1067.

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference (TCC)*, 2006, pp. 265–284.

[12] M. Abadi, A. Chu, I. Goodfellow, and *et al.*, "Deep learning with differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318.

[13] C. Chen and J. Lee, "Stochastic adaptive line search for differentially private optimization," in *IEEE International Conference on Big Data*, 2020, pp. 1011–1020.

[14] N. Papernot, A. Thakurta, S. Song, and *et al.*, "Tempered sigmoid activations for deep learning with differential privacy," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 9312–9321.

[15] M. Nasr, R. Shokri, and *et al.*, "Improving deep learning with differential privacy using gradient encoding and denoising," *arXiv:2007.11524*, 2020. [Online]. Available: https://arxiv.org/abs/2007.11524

[16] X. Chen, S. Z. Wu, and M. Hong, "Understanding gradient clipping in private sgd: A geometric perspective," in *Advances in Neural Information Processing Systems*, 2020, pp. 13 773–13 782.

[17] D. Yu, H. Zhang, W. Chen, and T. Liu, "Do not let privacy overbill utility: Gradient embedding perturbation for private learning," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=7aogOj_VYO0

[18] B. Wang, F. Wu, Y. Long, and *et al.*, "Datalens: Scalable privacy preserving training via gradient compression and aggregation," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021, pp. 2146–2168.

[19] N. Papernot, M. Abadi, U. Erlingsson, and *et al.*, "Semi-supervised knowledge transfer for deep learning from private training data," in *International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=HkwoSDPgg

[20] J. Hamm, Y. Cao, and M. Belkin, "Learning privately from multiparty data," in *International Conference on Machine Learning (ICML)*, 2016, pp. 555–563.

[21] N. Papernot, S. Song, I. Mironov, and *et al.*, "Scalable private learning with pate," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/pdf?id=rkZB1XbRZ

[22] Y. Zhu, X. Yu, M. Chandraker, and *et al.*, "Private-knn: Practical differential privacy for computer vision," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 854–11 862.

[23] L. Xie, K. Lin, S. Wang, and *et al.*, "Differentially private generative adversarial network," *arXiv:1802.06739*, 2018. [Online]. Available: https://arxiv.org/abs/1802.06739

[24] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=S1zk9iRqF7

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6114

[26] B. Liu, M. Ding, S. Shaham, and *et al.*, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1–36, 2022.

[27] V. M. Suriyakumar, N. Papernot, A. Goldenberg, and M. Ghassemi, "Chasing your long tails: Differentially private prediction in health care settings," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 723–734.

[28] M. Pathak, S. Rane, W. Sun, and *et al.*, "Privacy preserving probabilistic inference with hidden markov models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5868–5871.

[29] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2014, pp. 464–473.

[30] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 1310–1321.

[31] L. Xiang, L. Wang, S. Wang, and B. Li, "Achieving consensus in privacy-preserving decentralized learning," in *IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2020, pp. 899–909.

[32] Y. Gao, Q. Li, Y. Zheng, G. Wang, J. Wei, and M. Su, "Sedml: Securely and efficiently harnessing distributed knowledge in machine learning," *Computers & Security*, vol. 121, p. 102857, 2022.

[33] T. Miyato, S.-i. Maeda, M. Koyama, and *et al.*, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 8, pp. 1979–1993, 2018.

[34] G. Jagannathan, C. Monteleoni, and K. Pillaipakkamnatt, "A semi-supervised learning approach to differential privacy," in *IEEE International Conference on Data Mining Workshops*, 2013, pp. 841–848.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, and *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[36] Z. Dai, Z. Yang, F. Yang, and *et al.*, "Good semi-supervised learning that requires a bad gan," in *Advances in Neural Information Processing Systems*, 2017, pp. 6510–6520.

[37] V. Dumoulin, I. Belghazi, B. Poole, and *et al.*, "Adversarially learned inference," in *International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=B1ElR4cgg

[38] T. Salimans, I. Goodfellow, W. Zaremba, and *et al.*, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[39] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with gans: Manifold invariance with improved inference," in *Advances in Neural Information Processing Systems*, 2017, pp. 5540–5550.

[40] Y. Luo, J. Zhu, M. Li, and *et al.*, "Smooth neighbors on teacher graphs for semi-supervised learning," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8896–8905.

[41] G.-J. Qi, L. Zhang, H. Hu, and *et al.*, "Global versus localized generative adversarial nets," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1517–1525.

[42] B. Beaulieu-Jones, Z. Wu, C. Williams, and *et al.*, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 7, pp. 1–10, 2019.

[43] R. Torkzadehmahani, P. Kairouz, and B. Paten, "Dp-cgan: Differentially private synthetic data and label generation," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 98–104.

[44] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model," *arXiv:1801.01594*, 2018. [Online]. Available: https://arxiv.org/abs/1801.01594

[45] D. Chen, T. Orekondy, and M. Fritz, "Gs-wgan: A gradient-sanitized approach for learning differentially private generators," *Advances in Neural Information Processing Systems*, pp. 12 673–12 684, 2020.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.

[47] I. Gulrajani, F. Ahmed, M. Arjovsky, and *et al.*, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5769–5779.

[48] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006, pp. 535–541.

[49] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Advances in Neural Information Processing Systems Workshop on Deep Learning*, 2014, pp. 1–9.

[50] J. Gou, B. Yu, S. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision (IJCV)*, vol. 129, no. 6, pp. 1789–1819, 2021.

[51] A. Romero, N. Ballas, S. Kahou, and *et al.*, "Fitnets: Hints for thin deep nets," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6550

[52] Y. Liu, J. Cao, B. Li, and *et al.*, "Knowledge distillation via instance relationship graph," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7096–7104.

[53] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3967–3976.

[54] C. Xue, J. Yan, R. Yan, and *et al.*, "Transferable automl by model sharing over grouped datasets," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9002–9011.

[55] Z. Shen, Z. He, and X. Xue, "Meal: Multi-model ensemble via adversarial learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 4886–4893.

[56] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=SkgpBJrtvS

[57] S. Ge, S. Zhao, C. Li, and *et al.*, "Efficient low-resolution face recognition via bridge distillation," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 6898–6908, 2020.

[58] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 5573–5588, 2021.

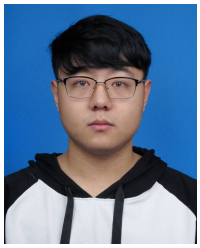
[59] Y. Feng, X. Sun, W. Diao, and *et al.*, "Double similarity distillation for semantic image segmentation," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 5363–5376, 2021.

[60] H. Lin, Y. Li, X. Fu, and *et al.*, "Rain o'er me: Synthesizing real rain to derain with data distillation," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 7668–7680, 2020.

[61] J. Wang, C. Hsieh, M. Wang, and *et al.*, "Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 914–922, 2021.

[62] H. Chen, Y. Wang, C. Xu, and *et al.*, "Data-free learning of student networks," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3514–3522.
[63] T. G. Dietterich, "Ensemble methods in machine learning," in *International Workshop on Multiple Classifier Systems*, 2000, pp. 1–15.
[64] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv:1606.01583*, 2016. [Online]. Available: https://arxiv.org/abs/1606.01583
[65] B. Yu, J. Wu, J. Ma, and *et al.*, "Tangent-normal adversarial regularization for semi-supervised learning," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 676–10 684.
[66] Y. Long, S. Lin, Z. Yang, and *et al.*, "Scalable differentially private generative student model via pate," *arXiv:1906.09338*, 2019. [Online]. Available: https://arxiv.org/abs/1906.09338
[67] F. Harder, K. Adamczewski, and M. Park, "Dp-merf: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021, pp. 1819–1827.
[68] P. Kairouz, H. B. McMahan, B. Avent, and *et al.*, "Advances and open problems in federated learning," *arxiv:1912.04977*, 2019. [Online]. Available: https://arxiv.org/abs/1912.04977
[69] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, pp. 211–407, 2014.
[70] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[71] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv:1708.07747*, 2017. [Online]. Available: https://arxiv.org/abs/1708.07747
[72] Y. Netzer, T. Wang, A. Coates, and *et al.*, "Reading digits in natural images with unsupervised feature learning," in *Advances in Neural Information Processing Systems Workshop on Deep Learning & Unsupervised Feature Learning*, 2011, pp. 1–9.
[73] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.
[74] L. Yu, L. Liu, C. Pu, and *et al.*, "Differentially private model publishing for deep learning," in *IEEE Symposium on Security and Privacy*, 2019, pp. 332–349.
[75] D. Yu, H. Zhang, W. Chen, and *et al.*, "Large scale private learning via low-rank reparametrization," in *International Conference on Machine Learning (ICML)*, 2021, pp. 12 208–12 218.
[76] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International Conference on Machine Learning (ICML)*, 2017, pp. 2642–2651.
[77] X. Chen, Y. Duan, R. Houthooft, and *et al.*, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2016, pp. 2180–2188.
[78] S. Warnat-Herresthal, H. Schultze, K. L. Shastry, and *et al.*, "Swarm learning for decentralized and confidential clinical machine learning," *Nature*, vol. 594, pp. 265–270, 2021.

Shiming Ge (M'13-SM'15) is a professor with the Institute of Information Engineering, Chinese Academy of Sciences. Prior to that, he was a senior researcher and project manager in Shanda Innovations, a researcher in Samsung Electronics and Nokia Research Center. He received the B.S. and Ph.D degrees both in Electronic Engineering from the University of Science and Technology of China (USTC) in 2003 and 2008, respectively. His research mainly focuses on computer vision, data analysis, machine learning and AI security, especially trustworthy learning solutions towards scalable applications. He is a senior member of IEEE, CSIG and CCF.



Bochao Liu received his B.S. degree in Electronical Information Science and Technology from the School of Information Science and Engineering in Shandong University, China. He is now a Ph.D Candidate at the Institute of Information Engineering at Chinese Academy of Sciences and the School of Cyber Security at the University of Chinese Academy of Sciences, Beijing. His major research interests are private-privacy machine learning.

Pengju Wang is an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He received the B.S. degree from the School of Information Science and Engineering at Shandong University and M.S. degree from the School of Electronic Engineering at Beijing University of Posts and Telecommunications. His research interests include AI security and federated learning.

Yong Li is an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He received the B.S. degree from the School of Computer Sciences at the Beijing Jiaotong University and Ph.D degree from the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include security data analysis and the design of private machine learning methods and systems.

Dan Zeng (SM'21) received her Ph.D. degree in circuits and systems, and her B.S. degree in electronic science and technology, both from University of Science and Technology of China, Hefei. She is a full professor and the Dean of the Department of Communication Engineering at Shanghai University, directing the Computer Vision and Pattern Recognition Lab. Her main research interests include computer vision, multimedia analysis, and machine learning. She is serving as the Associate Editor of the IEEE Transactions on Multimedia and the IEEE Transactions on Circuits and Systems for Video Technology, the TC Member of IEEE MSA and Associate TC member of IEEE MMSP.