Privacy-Preserving Student Learning with Differentially Private Data-Free Distillation

Bochao Liu*†, Jianghu Lu*†, Pengju Wang*†, Junjie Zhang‡, Dan Zeng‡, Zhenxing Qian§ and Shiming Ge*†
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
‡School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
§School of Computer Science, Fudan University, Shanghai 200433, China

***Abstract*—Deep learning models can achieve high inference accuracy by extracting rich knowledge from massive well-annotated data, but may pose the risk of data privacy leakage in practical deployment. In this paper, we present an effective teacher-student learning approach to train privacy-preserving deep learning models via differentially private data-free distillation. The main idea is generating synthetic data to learn a student that can mimic the ability of a teacher well-trained on private data. In the approach, a generator is first pretrained in a data-free manner by incorporating the teacher as a fixed discriminator. With the generator, massive synthetic data can be generated for model training without exposing data privacy. Then, the synthetic data is fed into the teacher to generate private labels. Towards this end, we propose a label differential privacy algorithm termed selective randomized response to protect the label information. Finally, a student is trained on the synthetic data with the supervision of private labels. In this way, both data privacy and label privacy are well protected in a unified framework, leading to privacy-preserving models. Extensive experiments and analysis clearly demonstrate the effectiveness of our approach.**

***Index Terms*—differential privacy, teacher-student learning, knowledge distillation**

I. Introduction

Deep learning models have proven success in many inference tasks [1]–[5] by extracting rich knowledge from massive well-annotated data. However, the deployment of these well-performing models may has risk of data privacy leakage since the training data often contain private information [6] and the models may be attacked. For example, the existing works [7], [8] have shown that the private information in the training data can be obtained from models even if the parameters are not known. Thus, it is important to devise effective solutions that can *learn privacy-preserving models with less accuracy loss*.

Differential privacy [9] is widely used in privacy protection and provides privacy measurement standard for data and label. Abadi *et al.* [10] first introduced differential privacy into stochastic gradient descent to train deep learning models, and their proposed DPSGD algorithm achieves good differential privacy but leads to large accuracy degradation. Papernot *et al.* [6] proposed private aggregation of teacher ensembles (PATE) that achieves differential privacy by limiting privacy loss with the number of labels. For only label-sensitive setting, Chaudhuri *et al.* [17] proposed the concept of label differential privacy (LabelDP). Badih *et al.* [11] then introduced prior probabilities into the randomized response and trained privacy-preserving models in a multi-step manner. Malek *et al.* [12] applied PATE and bayesian inference to the label differential privacy setting. Correspondingly, Yuan *et al.* [13] applied [10] to the label differential privacy setting. Esfandiari *et al.* [14] improved the model performance by adding a clustering operation before the randomized response. We found that it is much easier to perform the differential privacy algorithm on only the labels than on the data at the same time, and the model accuracy will be much higher through the above works. However, a major issue is how to convert differential privacy setting into label differential privacy setting.

To convert differential privacy setting into label differential privacy setting, a key is learning models with generative data and private labels. Recent data-free knowledge distillation can provide this function. Data-free knowledge distillation is a class of approaches which aims to train a student model with a pre-trained teacher model without access to original training data. It uses the information extracted from the teacher model to synthesize data used in the distillation process. Chen *et al.* [15] proposed data-free learning for training the student model by exploiting GAN. It uses the teacher model as a fixed discriminator to train a generator to generate the training data used for distillation. Fang *et al.* [16] proposed a FastDFKD that applied the idea of meta-learning to the training process to accelerate the efficiency of data synthesis. We found that a slight modification of the generator training method for such methods can learn only the data distribution information and ignore the data representation information.

Inspired by the above works, we propose a privacy-preserving data-free distillation method. As shown in Fig. 1, publishing a model (*e.g.*, teacher model) trained directly from private data would compromise privacy, so we treat it as a fixed discriminator to train a generator in a data-free manner. This generator learns only the data distribution to protect the private data. Using this generator implicitly generates data for the distillation process from teacher model to student model. Because querying the teacher model using the generated synthetic data can compromise private information, we propose a LabelDP algorithm selective randomized response to protect the output of the teacher model. The selective randomized

Shiming Ge is the corresponding author. Email: geshiming@iie.ac.cn.

978-1-6654-7189-3/22/$31.00 © 2022 IEEE

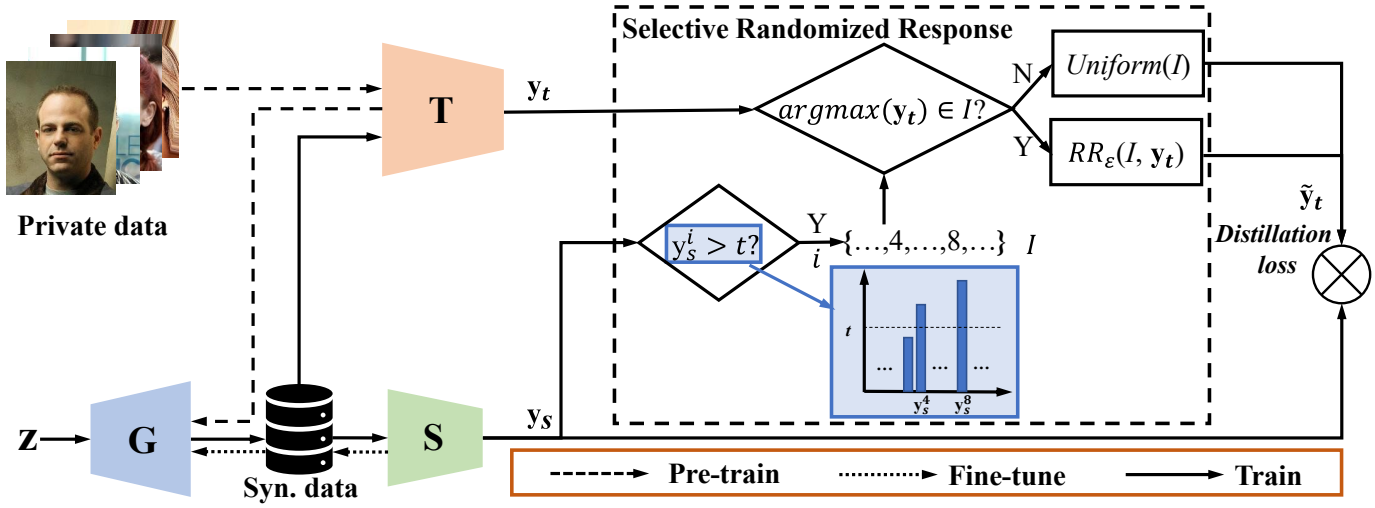

Fig. 1. The framework of our differentially private data-free distillation approach. It aims to train a privacy-preserving student model **S** with teacher-student learning. First, a teacher **T** is well trained on private data and serves as a fixed discriminator to pre-train a generator **G** in a data-free manner. Then, massive synthetic data is generated from noisy code **z** with the generator and fed into the teacher and student **S** to query differentially private labels with selective randomized response. Finally, with the synthetic data and noisy labels, the student is trained by regressing the teacher knowledge. In this way, both the data privacy and label privacy are well protected in a unified framework, leading to a privacy-preserving student model **S** doing the distillation with final labels and outputs of student. In the selective randomized response, we use the output of the student model combined with a threshold t to reduce the number of possible labels and obtain I. We implement ε-DP with return $RR_\varepsilon(I, \mathbf{y}_t)$ if correct label in I and $Uniform(I)$ if correct label not in I.

response algorithm treats the output of the student model as prior knowledge to reduce the possible output categories to increase the probability of outputting the correct label, and if the possible output does not contain the correct label, a uniform probability distribution is used to reset the possible probability of the output. In summary, our approach can effectively learn privacy-preserving student model by two keys. On the one hand, our proposed data-free distillation is able to protect privacy well with the learning of data distribution. The generated synthetic data from this generator will not reveal private information even if it is distributed. On the other hand is that we propose the selective randomized response module to implement DP, which is no longer limited by the number of queries, and introduce the prediction of the student model as prior knowledge for the randomized response. We increase the probability of returning the correct label by setting a threshold, so the student model can learn the knowledge of the teacher model more effectively.

Our major contributions are three folds: 1) we propose a differentially private data-free distillation approach to learn privacy-preserving and high accurate student models via synthetic data, 2) we propose selective randomized response algorithm to privately distill teacher knowledge which provides strong protect label privacy protection in theory, and 3) we conduct extensive experiments and privacy analysis to demonstrate the effectiveness of our approach.

II. Approach

A. Problem Formulation

Given a private dataset $\mathcal{D}$, the goal is to train a student model ϕ_s with privacy-preserving capabilities and its accuracy close to the teacher model ϕ_t trained directly on $\mathcal{D}$. To achieve this goal, we propose a privacy-preserving differentially private data-free distillation method. First, we train a teacher model ϕ_t directly on $\mathcal{D}$. Then, we use ϕ_t as a fixed discriminator to train a generator ϕ_g that is used to generate massive synthetic data $\tilde{\mathcal{D}}$. We obtain predictions on $\tilde{\mathcal{D}}$ by querying the teacher model and apply the selective randomized response function which follows ε-LabelDP on them to get labels $\mathcal{L}$. Finally, the student learning can be formulated by minimizing an energy function $\mathbb{E}$:

$$\mathbb{E}(\theta_s; \tilde{\mathcal{D}}) = \mathbb{E}(\phi_s(\theta_s; \tilde{\mathcal{D}}), \mathcal{L}) = \mathbb{E}(\phi_s(\theta_s; \tilde{\mathcal{D}}), \mathcal{R}(\phi_t(\theta_t; \tilde{\mathcal{D}}))), \tag{1}$$

where θ_s and θ_t are the parameters of the student and teacher, respectively. $\mathcal{R}$ is selective randomized response function.

From Eq. 1, we can see that our approach can learn privacy-preserving models by two main processes. First, the training of student does not directly access the private data. Second, the labels from the teacher are protected by the selective randomized response module which implements ε-LabelDP. Therefore, privacy leakage can be suppressed very effectively. During training, the teacher knowledge is transferred to the student through the label $\mathcal{L}$. We solve Eq. 1 via two steps, including: 1) data-free generator learning that trains a generator ϕ_g with the pre-trained teacher ϕ_t as a fixed discriminator to generate synthetic data $\tilde{\mathcal{D}}$, and 2) student learning that applies knowledge distillation is to label the synthetic data with ϕ_t and selective randomized response function. And then use these data-label pairs to train the student model ϕ_s and fine-tune the generator ϕ_g. The detailed process is introduced in Alg. 1.

B. Data-Free Generator Learning

Directly using private data to train the generator will lead to privacy leakage, while using public data will lead to a serious

Algorithm 1: Differentially Private Data-Free Distillation (DP-DFD)

Input: Number of stages T; Pre-trained teacher ϕ_t; Initial student $\phi_s^{(0)}$; Threshold t.

Training a generator ϕ_g with the teacher model ϕ_t as a fixed discriminator in data-free manner.

For $i = 1$ to T:

a) Generate synthetic data $\tilde{\mathcal{D}}^{(i)}$;

b) Compute $\mathbf{y}_t$ and $\mathbf{y}_s$ by entering $\tilde{\mathcal{D}}^{(i)}$ into ϕ_t and $\phi_s^{(i-1)}$;

c) Let $\hat{\mathcal{D}}^{(i)} = [\tilde{\mathcal{D}}^{(i)}, \mathcal{R}(\mathbf{y}_s, \mathbf{y}_t, t)]$;

d) Train the student model $\phi_s^{(i)}$ and fine-tune ϕ_g on $\hat{\mathcal{D}}^{(i)}$.

Return $\phi_s^{(T)}$.

Function selective randomized response $\mathcal{R}$:

Input: Output of student model $\mathbf{y}_s$; The teacher output y_t; Threshold t;

Select the set I of indexes with condition $y_s^i > t$, and $k = |I|$ is the set size.

if $|I| = 0$ *or* $|I| = 1$ **then**

  I = [index of top two largest elements in y_s]

if $\arg\max(\mathbf{y}_t) \in I$ **then**

  Return $\mathbf{y}_t$ with probability $\frac{e^\varepsilon}{e^\varepsilon+k-1}$ and the one-hot type of other elements with probability $\frac{1}{e^\varepsilon+k-1}$

else

  Return the one-hot type of all elements in I with probability $\frac{1}{k}$

decrease in the accuracy of the student model obtained by distillation, so we want to find a generator training method that does not leak privacy and could match the distribution of the private data. Inspired by [15], multi-class classifiers instead of two-class classifiers as discriminators can better learn data distribution, so we adopt a new training approach. We first train a teacher model directly using the private data, and then train a generator using that teacher model as a discriminator with fixed parameters. At the heart of this idea is to take the teacher as a bridge to indirectly learn the distribution of private data. We optimize them by the following loss:

$$\begin{aligned}\mathcal{L}_g(\tilde{\mathbf{x}}) =& \ell_{CE}(\phi_t(\theta_t;\tilde{\mathbf{x}}), \arg\max_j(\phi_t(\theta_t;\tilde{\mathbf{x}}))_j) + \\ & \alpha\phi_t(\theta_t;\tilde{\mathbf{x}})\log\phi_t(\theta_t;\tilde{\mathbf{x}}) + \beta\mathcal{N}(\phi_t,\tilde{\mathbf{x}}),\end{aligned} \tag{2}$$

where $\tilde{\mathbf{x}} = \phi_g(\theta_g;\mathbf{z})$ generated by ϕ_g with parameters θ_g, $\mathbf{z}$ is a random vector, α and β are the tuning parameters to balance the effect of three terms. The cross entropy $\ell_{CE}(\cdot)$ is used to enforce the outputs of the teacher model closer to the one-hot labels. The smaller it is, the closer the synthetic data distribution is to the private data. The second term is the information entropy loss to measure the class balance of synthetic data. The $\mathcal{N}(\cdot)$ is l_2-norm $||*||_2$ to measure the mean and variance of the total synthetic data and the running data fed into the model. In this way, the synthetic data $\tilde{\mathcal{D}}$ generated by the trained generator has a similar distribution to private data without compromising privacy.

C. Student Learning with Synthetic Data and Private Labels

During the training of the student model, we use randomized response [18] for the sensitive labels to achieve DP. RR_ε mechanism will return correct class label with the probability $\frac{e^\varepsilon}{e^\varepsilon+K-1}$, and return other labels with probability $\frac{1}{e^\varepsilon+K-1}$, where K is the number of classes. To improve the probability of returning the true label without compromising privacy, we introduce the student prediction $\mathbf{y}_s$ and propose selective randomized response algorithm. As shown in function selective randomized response in Alg. 1, we first set a threshold t and select the set of indexes I with condition $\mathrm{y}_s^i > t$. To ensure the randomness of the output labels, we require that the number of elements in I to be at least 2. We will set I to the set of indexes of top two largest elements in $\mathbf{y}_s$ if the number of elements in I is less than 2. Let k be the number of I. If the teacher model's output in I, return the $\mathbf{y}_t$ with the probability $\frac{e^\varepsilon}{e^\varepsilon+k-1}$ and return the one-hot type of other elements with probability $\frac{1}{e^\varepsilon+k-1}$ ($RR_\varepsilon(I, \mathbf{y}_t)$ in Fig. 1). If the teacher model's output not in I, return the one-hot type of the elements in I with probability $\frac{1}{k}$ ($Uniform(I)$ in Fig. 1).

For learning with LabelDP guarantee, we use selective randomized response to randomized outputs from the teacher model for each example of the synthetic data and then apply a general learning algorithm that is robust to random label noise to these data-label pairs. Unlike DPSGD and PATE, which require the composition theorems to calculate the final privacy budget ε, we query the random labels once and reuse them in training process. At each stage $i \in [T]$, the synthetic dataset $\tilde{\mathcal{D}}^{(i)}$ is first generated using the generator ϕ_g and then enter it into the most recent student model $\phi_s^{(i-1)}$ to obtain $\mathbf{y}_s$ as the prior knowledge. We run selective randomized response algorithm with $\mathbf{y}_s$ to obtain the label $\mathcal{L}_i$. We use $\hat{\mathcal{D}}^{(i)} = \{\tilde{\mathcal{D}}^{(i)}, \mathcal{L}_i\}$ to train the student model $\phi_s^{(i)}$ and fine tune the generator ϕ_g. The loss function for the ith epoch is

$$\mathcal{L}_{kd}^{(i)} = \sum_{j=1}^{|\hat{\mathcal{D}}^{(i)}|} \ell_{KL}(\phi_s^{(i)}(\theta_s;\tilde{\mathbf{x}}_j), \tilde{\mathbf{y}}_j),\ s.t.\ (\tilde{\mathbf{x}}_j, \tilde{\mathbf{y}}_j) \in \hat{\mathcal{D}}^{(i)}, \tag{3}$$

where $\ell_{KL}(\cdot)$ represents the Kullback-Leibler divergence. For the synthetic dataset $\tilde{\mathcal{D}} = \cup_{i=1}^T \tilde{\mathcal{D}}^{(i)}$, iff the size of dataset in each stage is the same, their order will have no effect on the accuracy of student, but T will affect the student accuracy.

In our approach, the private data first transfers the knowledge to the teacher model. Directly publishing the teacher model would lead to privacy leakage, so we use the teacher model as a fixed discriminator to train a generator to generate a non-sensitive synthetic dataset with the similar distribution to the private data. We use this dataset to transfer the knowledge from the teacher model to the student model. Because only the predictions from the teacher model are sensitive, we protect the predictions of the teacher model by implementing ε-LabelDP through our propose selective randomized response module. In this way, the knowledge of the private data is transferred to the privacy-preserving student model without access through a data-free distillation approach.

III. EXPERIMENTS

To verify the effectiveness of our differentially private data-free distillation approach (**DP-DFD**), we conduct experiments on five datasets and perform comprehensible comparisons with 11 state-of-the-arts. To make the comparisons fair, our experiments use the same settings as these approaches and take the results from their original papers.

A. Experimental Setting

Datasets. The experiments are conducted on five datasets. MNIST [19] and FashionMNIST (FMNIST) [20] are both 10-class datasets for 28×28 gray handwritten number images and fashion images, respectively. They includes 60K train examples and 10K test examples. CIFAR10 and CIFAR100 [21]) consists of 60K 32×32 color object images in 10 and 100 subjects, where 50K for training and 10K for testing. CelebA [5] contains 202,599 color facial images that are preprocessed by aligning and resizing into 64×64. According to hair color and gender attributes, we create two CelebA datasets, CelebA-H and CelebA-G and they uses black/blonde/brown and male/female as labels, respectively. We partition them into training set and test set according to the official criteria [5].

Implementation. In all experiments, the structure of teacher is $Resnet34$ and we set α and β in Eq. 2 as 5 and 10 respectively. The structure of student is the same as [27] in data-sensitive experiments and $Resnet18$ in label-sensitive experiments, respectively. For each dataset, we set the threshold value to $1/(2 * nc)$ where nc is the number of classes. We evaluate the test accuracy of student under privacy protection.

B. State-of-the-art Comparisons

Comparisons with data-sensitive approaches. First, we compare with 7 data-sensitive approaches on MNIST, FMNIST, CelebA-H and CelebA-G under ε=1 and ε=10, including DP-GAN [22], PATE-GAN [23], GS-WGAN [24] and G-PATE [25], DP-MERF [26], DataLens [27] and DP-Sinkhorn [28]. The results are shown in Tab. I. All other approaches are under a failure probability $\delta = 10^{-5}$. The accuracy of baseline model which trained directly using private data is 99.21% on MNIST, 91.02% on FMNIST, 93.53% on CelebA-G and 88.68% on CelebA-H. From Tab. I, we can see that our DP-DFD shows substantially higher performance than other approaches especially when $\varepsilon = 1$. In particular, the accuracy of our DP-DFD outperforms the other best-performing approaches by 21 percentage points. When $\varepsilon = 10$, our DP-DFD also has an absolute advantage and is at least 13 percentage points higher than the other approaches. Even for high-dimensional datasets like CelebA-G and CelebA-H, our DP-DFD still shows the state-of-the-art performance, which also demonstrates the advantage of DP-DFD over other privacy-preserving approaches on high-dimensional datasets.

Comparisons with label-sensitive approaches. Then, we compare with 4 LabelDP approaches (LP-MST [11], AL-IBI [12], ClusterRR [14] and Protocol [13]) on MNIST, FMNIST, CIFAR10 and CIFAR100 under the same ε. The results are shown in Tab. II. We can find that our method performs optimally for all four datasets and for different ε. In particular, it achieves a correct rate of 74.67 when ε equals 8 on the CIFAR100 dataset, surpassing many methods that use the raw data for direct distillation. The effectiveness of our method is further demonstrated by the fact that our method has higher performance than the other four methods trained directly using raw data when there is no restriction on the amount of generated synthetic data.

C. Ablation Studies

After the promising performance is achieved, we further analyze each influencing factor in our approach, including the impact of loss terms in the data-free generator learning, the amount of synthetic data and the number of stages.

Loss function. To further understand the improvement of each component of the loss function during data-free training of the generator, we designed experiments on MNIST and FMNIST under ε=10 to explore the contribution of each component. The results are shown in Tab. III. where CE means the cross entropy loss term, IE is the information entropy loss term and Norm is the normalized term for the mean and variance of the data. We can see that the normalization term of the data has the greatest impact, followed by the information entry loss term and finally the cross entropy loss term. We speculate that this may be related to the randomness of the data generated by the generator, which limits the distribution of the data to make the generated synthetic data more usable, so it has a greater impact on the accuracy of the student model.

Data amount. We further conducted experiments on MNIST, FMNIST, CIAFR10 and CIFAR100 datasets under $\varepsilon = 1$. The results are shown in Fig. 2. We found that MNIST dataset converges at about 50,000 data volume, FMNIST converges at about 120,000, CIFAR10 and CIFAR100 converge at about 220,000 and 500,000, respectively. As the difficulty of datasets increases, the amount of data required to achieve convergence increases. We suspect that this is because the more difficult the dataset is, the more difficult its distribution knowledge is to learn, so the larger the amount of data required. We note that the CIFAR10 dataset is more difficult than FMNIST, but the reason why CIFAR10's final accuracy is similar to FMNIST's is that the network structure is different.

Number of stages. To explore the effect of the number of stages, we conducted experiments on MNIST, FMNIST and CIFAR10 datasets under ε=10. The results are shown in Fig. 3. Experimental results show that between 20 and 320, the accuracy of the student model increases with the increase of stages. As the classification difficulty of MNIST, FMNIST and CIFAR10 datasets increases, the effect of stages becomes greater. The experimental results are as we expected because we used the prediction of the student model as the prior knowledge. As the training process proceeds, the more accurate the prediction of the student model becomes, which means the higher the probability of outputting the correct label.

TABLE I
ACCURACY COMPARISONS WITH 7 DATA-SENSITIVE APPROACHES: TEST ACCURACY UNDER DIFFERENT PRIVACY BUDGET ε.

Dataset	Baseline	ε	DP-GAN	PATE-GAN	G-PATE	GS-WGAN	DP-MERF	DataLens	DP-Sinkhorn	DP-DFD
MNIST	**0.9921**	1	0.4036	0.4168	0.5810	0.1432	0.6500	0.7123	-	**0.9762**
		10	0.8011	0.6667	0.8092	0.8075	0.6870	0.8066	0.8320	**0.9856**
FMNIST	**0.9102**	1	0.1053	0.4222	0.5567	0.1661	0.6100	0.6478	-	**0.8917**
		10	0.6098	0.6218	0.6934	0.6579	0.6250	0.7061	0.7110	**0.9074**
CelebA-G	**0.9353**	1	0.5330	0.6068	0.6702	0.5901	-	0.7058	-	**0.7814**
		10	0.5211	0.6535	0.6897	0.6136	0.6500	0.7287	0.7630	**0.8934**
CelebA-H	**0.8868**	1	0.3447	0.3789	0.4985	0.4203	-	0.6061	-	**0.6753**
		10	0.3920	0.3900	0.6217	0.5225	-	0.6224	-	**0.8207**

TABLE II
ACCURACY COMPARISONS WITH 4 LABEL SENSITIVE APPROACHES: TEST ACCURACY UNDER DIFFERENT PRIVACY BUDGET ε.

Dataset	ε	LP-2ST	ALIBI	ClusterRR	Protocol	DP-DFD
MNIST	1	0.9582	-	0.9000	-	**0.9762**
FMNIST	1	0.8326	-	0.8800	-	**0.8917**
CIFAR10	1	0.6367	0.8420	0.6857	-	**0.8796**
	2	0.8605	-	-	0.8184	**0.8812**
CIFAR100	3	0.2874	0.5500	-	-	**0.5861**
	8	0.7410	0.7440	-	-	**0.7467**

TABLE III
IMPACT OF LOSS TERMS IN TRAINING GENERATOR UNDER ε=10.

Dataset	CE	IE	Norm	Accuracy
MNIST	✓	✓	✓	**0.9856**
	✗	✓	✓	0.9655
	✓	✗	✓	0.9432
	✓	✓	✗	0.8801

The greater the percentage of synthetic data being correctly labeled, the better the student model performance will be.

D. Privacy-Preserving Analysis

Data generation. To demonstrate that the direct use of synthetic data in our approach doesn't leak information of private data, we visualize some examples for MNIST, FMNIST, CIFAR10 and CelebA, as shown in Fig. 4. The first row is MNIST, followed by FMNIST, CIFAR10, CelebA-G and CelebA-H in that order. We found that even for the simplest MNIST synthetic data, we could not semantically identify it as a handwritten font. Despite its inability to be recognized by humans, it has high utility in terms of training high performance models. We also found something interesting: such synthetic data can train a model that performs well, which raises an interesting question about what machine learning models actually learn from data?

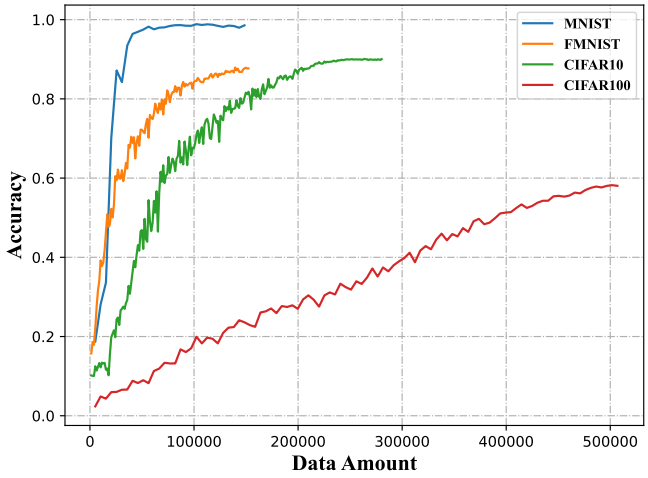

Fig. 2. The effect of different amount of synthetic data (ε=1).

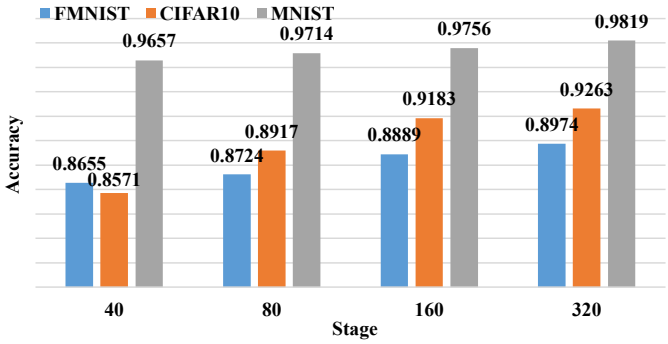

Fig. 3. The effect of different number of stages (ε=10).

Model-inversion attack. We perform a model-inversion attack [8] on a typical data-sensitive approach and a label-sensitive approach to further demonstrate that our approach can protect data privacy. The results are shown in Fig. 5. The first row is the results of the attack on a typical data-

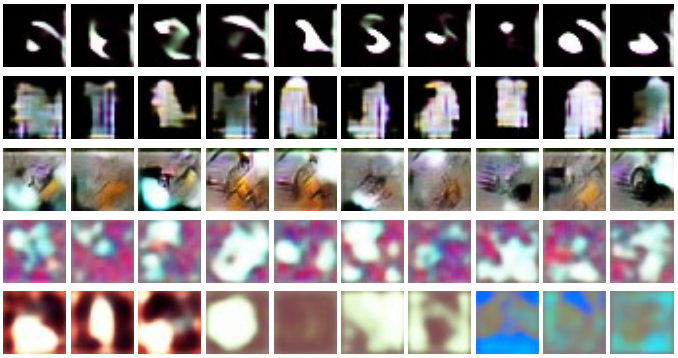

Fig. 4. The examples of the generated synthetic data. From top to bottom: MNIST, FMNIST, CIFAR10, CelebA-G and CelebA-H.

Fig. 5. The results of model-inversion attack against the students trained on MNIST with DataLens (top), ALIBI (middle) and DP-DFD (bottom).

sensitive method DataLens [27], while the second row shows the results of the attack on a typical label-sensitive method ALIBI [12]. The last row is the results of the attack on our DP-DFD. We emphasize that the authors of [8] stress in their original paper that differential privacy hardly works against this attack method, but we can find that even for experiments on the simplest MNIST dataset, our method still can defend against this attack and protect the privacy of the private data.

IV. Conclusion

Typically, publishing deep learning models may pose the risk of privacy leakage. To facilitate model deployment, we propose a differentially private data-free distillation approach (DP-DFD) that does not use private data in the training process of publish model. This approach uses the teacher model trained directly with private data as a bridge to transfer knowledge from private data to publish model. The generator trained in a data-free manner can learn the distribution of the private data and enhance the knowledge of the publish model to compensate for the loss of the accuracy without compromising privacy. In addition, we also provide differential privacy analysis for our selective randomized response and DP-DFD to demonstrate that it provides strong privacy guarantees in theory. We have conducted extensive experiments and analyses to show the effectiveness of our approach. In the future, we will explore the approach in more practical applications, such as federated learning on medical images and financial data.

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

References

[1] 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.
[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in NeurIPS, pp. 1106–1114, 2012.
[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, pp. 770–778, 2016.
[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR, 2020.
[5] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in ICCV, pp. 3730–3738, 2015.
[6] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in ICLR, 2017.
[7] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in ACM CCS, pp. 1322–1333, 2015.
[8] Y. Zhang, R. Jia, H. Pei, W. Wang, B Li and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in CVPR, pp. 253–261, 2020.
[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in TCC, pp. 265–284, 2006.
[10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, M. Ilya, T. Kunal and *et al.*, "Deep learning with differential privacy," in ACM CCS, pp. 308–318, 2016.
[11] B. Ghazi, N. Golowich, R. Kumar, P. Manurangsi and C. Zhang, "Deep learning with label differential privacy," in NeurIPS, pp. 27131-27145, 2021.
[12] M. Malek Esmaeili, I. Mironov, K. Prasad, I. Shilov, and F. Tramer, "Antipodes of label differential privacy: Pate and alibi," in NeurIPS, pp. 6934-6945, 2021.
[13] S. Yuan, M. Shen, I. Mironov, and A. Nascimento, "Label private deep learning training based on secure multiparty computation and differential privacy," in NeurIPS Workshop, 2021.
[14] H. Esfandiari, V. Mirrokni, U. Syed, and S. Vassilvitskii, "Label differential privacy via clustering," in AISTATS, pp. 7055–7075, 2022.
[15] H. Chen, Y. Wang, C. Xu, Z. Yang and C. Liu, "Data-free learning of student networks," in ICCV, pp. 3514–3522, 2019.
[16] G. Fang, K. Mo, X. Wang, J. Song, S. Bei, H. Zhang and M. Song, "Up to 100x faster data-free knowledge distillation," in AAAI, 2022.
[17] K. Chaudhuri and D. Hsu, "Sample complexity bounds for differentially private learning," in COLT, pp. 155-186, 2011.
[18] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," JASA, vol. 60, no. 309, pp. 63–69, 1965.
[19] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[20] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," arXiv:1708.07747, 2017.
[21] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.
[22] L. Xie, K. Lin, S. Wang, F. Wang and J. Zhou, "Differentially private generative adversarial network," arXiv:1802.06739, 2018.
[23] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in ICLR, 2019.
[24] D. Chen, T. Orekondy, and M. Fritz, "Gs-wgan: A gradient-sanitized approach for learning differentially private generators," in NeurIPS, pp. 12:673–12:684, 2020.
[25] Y. Long, S. Lin, Z. Yang, C. A. Gunter and B. Li, "Scalable differentially private generative student model via PATE," arXiv:1906.09338, 2019.
[26] F. Harder, K. Adamczewski, and M. Park, "Dp-merf: Differentially private mean embeddings with random features for practical privacy preserving data generation," in AISTATS, pp. 1819–1827, 2021.
[27] B. Wang, F. Wu, Y. Long, L. Rimanic and C. Zhang, "Datalens: Scalable privacy preserving training via gradient compression and aggregation," in ACM CCS, pp. 2146–2168, 2021.
[28] T. Cao, A. Bie, A. Vahdat, S. Fidler and K. Kreis, "Don't generate me: Training differentially private generative models with sinkhorn divergence," in NeurIPS, pp. 12480-12492, 2021.